

Gunther Josef Schaubberger

Regularization Methods for Item Response and Paired Comparison Models

Dissertation an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

Eingereicht am 01.10.2015

Gunther Josef Schaubberger

Regularization Methods for Item Response and Paired Comparison Models

Dissertation an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

Eingereicht am 01.10.2015

Erster Berichterstatter: Prof. Dr. Gerhard Tutz
Zweiter Berichterstatter: Prof. Dr. Helga Wagner

Tag der Disputation: 26.11.2015

Danksagung

Ich möchte mich bei allen bedanken, die zur Entstehung dieser Doktorarbeit beigetragen haben. Insbesondere möchte ich mich bedanken bei ...

- ... meinem Doktorvater Gerhard Tutz für die hervorragende Betreuung, die angenehme Arbeitsatmosphäre, den stets herzlichen und humorvollen Umgang und viele konstruktive Diskussionen.
- ... Helga Wagner, die sich ohne zu Zögern bereit erklärt hat als Zweitgutachterin für diese Arbeit zur Verfügung zu stehen.
- ... Christian Heumann als Vorsitzendem der Prüfungskommission und besonders für die gute Zusammenarbeit in der Lehre.
- ... meinen ehemaligen und aktuellen Kollegen des Seminars für angewandte Stochastik für offene Türen, diverse Institutzkolloquien und eine stets angenehme Arbeitsatmosphäre.
- ... allen am Institut für Statistik.
- ... der Statistical Modelling Society für das jährliche Ausrichten des IWSM und bei allen, die diese Konferenz zu etwas Besonderem machen.
- ... der Leibspeiserei für Gemüselasagne, Bulgursalat, Pasta, Pasta, Pasta, ...
- ... Bastian Schweinsteiger und der ganzen Nationalmannschaft für den Sieg im WM-Finale, nicht nur im Hinblick auf das WM-Paper.
- ... meinen Freunden und Kollegen innerhalb des Instituts für zahlreiche Mittagessen, Kaffeepausen und noch mehr gemeinsame Aktivitäten außerhalb der Arbeit. Ganz besonders bedanken möchte ich bei Linda, Tina und Verena.
- ... meinen Freunden außerhalb des Instituts, insbesondere bei Holger, Michael und Ingrid für viele schöne Reisen und alles Andere.
- ... meiner Familie, besonders bei meinen Eltern, meinen Schwestern sowie bei Pia, Franz-Xaver und Miriam für Ablenkung, Rückhalt und die Besinnung auf das Wesentliche im Leben.

Zusammenfassung

Ein Hauptaspekt der psychometrischen Modellierung liegt in der Messung latenter Eigenschaften. Variablen oder Eigenschaften werden als latent bezeichnet wenn sie nicht direkt messbar sind und durch andere, direkt beobachtbare, Variablen ersetzt werden müssen.

Item Response Daten dienen dazu, nicht unmittelbar beobachtbare Fähigkeiten oder Eigenschaften zu messen. Bei Intelligenztests beispielsweise muss jeder Teilnehmer versuchen eine Reihe von Aufgaben zu lösen die so gestaltet sind, dass man mit ihnen die latente Eigenschaft Intelligenz messen kann. Die gängigste Art Item Response Daten zu modellieren ist das Rasch Modell. Es verwendet einen Parameter für die Fähigkeit der Person und einen Parameter für die Schwierigkeit der Aufgabe um die Wahrscheinlichkeit zu modellieren, dass eine bestimmte Person eine bestimmte Aufgabe löst.

Die Messung latenter Eigenschaften ist auch das Ziel von Paarvergleichen, die zustande kommen wenn zwei Objekte bezüglich bestimmter Eigenschaften verglichen werden. Beispielsweise können sportliche Wettkämpfe zwischen zwei Kontrahenten als Paarvergleiche bezüglich der Fähigkeiten der Kontrahenten gesehen werden. Außerdem können Paarvergleiche in experimentellen Designs verwendet werden um unbeobachtbare Eigenschaften zu messen, beispielsweise die Attraktivität verschiedener Produkte. Das gängigste Paarvergleichsmodell ist das sogenannte Bradley-Terry-Luce Modell. Es modelliert die Wahrscheinlichkeit, dass ein Objekt ein anderes übertrifft oder einem anderen Objekt gegenüber bevorzugt wird mit Hilfe der Differenz zwischen den geschätzten Eigenschaften beider Objekte.

Sowohl das Rasch Modell als auch das Bradley-Terry-Luce Modell können in das Konzept der Generalisierten Linearen Modelle eingebettet werden. Innerhalb dieses Konzepts sind zahlreiche Erweiterungen möglich. In dieser Arbeit werden beide Modelle durch das Einbeziehen von Kovariablen erweitert, was zu allgemeineren Modellen führt. Das Einbeziehen von Kovariablen eröffnet neue Einblicke in die Strukturen latenter Eigenschaften. Insbesondere können die recht starken Annahmen, die sowohl das Rasch Modell als auch das Bradley-Terry-Luce Modell unterstellen, abgeschwächt werden. Die größere Flexibilität der vorgeschlagenen Methoden führt aber auch zu einer größeren Komplexität der Modelle. Regularisierungsmethoden erweisen sich als probates Instrument um mit der größeren Anzahl an Parametern umzugehen und um zwischen notwendigen und unnötigen Parametern zu unterscheiden. Die vorgeschlagenen Methoden werden anhand von verschiedenen Simulationen und Anwendungen auf echte Daten veranschaulicht.

Im Rasch Modell sind die Kovariablen notwendig, um Aufgaben mit sogenanntem Differential Item Functioning (DIF) ausfindig zu machen. Differential Item Functioning tritt dann auf, wenn die Wahrscheinlichkeit, eine Aufgabe zu lösen, sich für zwei Personen mit der

gleichen Fähigkeit unterscheidet. Die vorgeschlagene Erweiterung des Rasch Modells erlaubt es, dass die Aufgabenschwierigkeiten von personenspezifischen Kovariablen abhängen und kann dadurch Aufgaben mit Differential Item Functioning ermitteln. Zur Schätzung wurden verschiedene Methoden entwickelt: DIFlasso verwendet einen Penalisierungsansatz während DIFboost auf Boosting Strategien basiert. Im Gegensatz zu den meisten existierenden Methoden um Differential Item Functioning aufzudecken ermöglicht es die Methode sowohl kategoriale als auch stetige Kovariablen sowie mehrere Kovariablen gleichzeitig zu verwenden.

Im Bradley-Terry-Luce Modell muss man zwischen objektspezifischen und subjektspezifischen Kovariablen unterscheiden. Sowohl Attribute des Subjekts, das die Entscheidung trifft, als auch Attribute der entsprechenden Objekte können möglicherweise die Entscheidung zwischen den Objekten beeinflussen. Unterschiedliche neue Modellierungsansätze und Penalisierungsterme werden diskutiert die dazu geeignet sind, mit den Anforderungen umzugehen, die sich aus den jeweiligen Typen von Kovariablen ergeben.

Summary

A main aspect in psychometric modeling is the measurement of latent traits. Variables or traits are called latent if they can not be measured directly and need to be replaced by other, directly observable, variables. This dissertation focuses on two popular methods to analyze latent traits, namely item response methods and paired comparisons.

Item response data serve to measure not directly observable abilities or traits. For example, in intelligence tests every participant faces a series of items which are designed to measure the latent trait intelligence. The most popular way to model item response data is the Rasch model. It uses one parameter for the ability of the person and one for the difficulty of the item to model the probability that a specific person solves a specific item.

The measurement of latent traits is also the goal of paired comparisons which appear when two objects are compared with respect to specific traits. For example, sport competitions between two opponents can be seen as paired comparisons with respect to the abilities of the opponents. Furthermore, paired comparisons can be used in experimental designs to measure unobservable traits, for example the attractiveness of different products. The most popular model for paired comparisons is the so-called Bradley-Terry-Luce model. It models the probability that one object beats (or is preferred over) another object using the difference between the estimated traits of both objects.

Both the Rasch model and the Bradley-Terry-Luce model can be embedded into the framework of generalized linear models. Within the framework of generalized linear models, several extensions are possible. In this thesis, both models are extended by the incorporation of covariates leading to more general models. The inclusion of covariates allows for new insights into the structure of the latent traits. In particular, it allows to weaken the rather strong assumptions implied by the Rasch model and the Bradley-Terry-Luce model. The increased flexibility of the proposed models also leads to a higher complexity of the models. Regularization methods prove to be an effective instrument to deal with the increased number of parameters and to differentiate between necessary and unnecessary parameters. The proposed methods are illustrated in various simulations and real data applications.

In the Rasch model, the covariates are essential to identify items with differential item functioning (DIF). Differential item functioning appears if the probability to solve an item is different for persons with the same ability. The proposed extension of the Rasch model allows for item difficulties to depend on person-specific covariates and, therefore, the new model can identify items with differential item functioning. Different estimation methods for the new model are developed: DIFlasso uses a penalty approach while DIFboost is based on boosting strategies. In contrast to most existing methods to detect differential item functioning, the methods allow to use both categorical and continuous covariates and to use several covariates simultaneously.

In the Bradley-Terry-Luce model, one has to distinguish between object-specific covariates and subject-specific covariates. Both attributes of the subject that decides and attributes of the respective objects can possibly affect the decision between the objects. Different new modeling approaches and penalty terms are discussed which are suited to deal with the challenges caused by the respective types of covariates.

Contents

1. Introduction	1
2. The Rasch Model	9
2.1. Assumptions and Properties of the Rasch Model	10
2.2. Estimation Approaches for the Rasch Model	11
3. Differential Item Functioning	15
3.1. Popular Methods for DIF Detection	16
3.2. Problems and Limitations	17
4. A Penalty Approach to Differential Item Functioning in Rasch Models	19
4.1. Introduction	19
4.2. Differential Item Functioning Model	20
4.2.1. The Binary Rasch Model	20
4.2.2. A General Differential Item Functioning Model	21
4.3. Estimation by Regularization	24
4.3.1. Maximum Likelihood Estimation	24
4.3.2. Penalized Estimation	25
4.4. The Fitting Procedure At Work	29
4.5. Examples	38
4.5.1. Exam Data	38
4.5.2. Knowledge Data	40
4.6. DIFlasso with Variable Selection	42
4.7. An Alternative Method	44
4.8. Concluding Remarks	46
5. Detection of Differential Item Functioning in Rasch Models by Boosting Techniques	47
5.1. Introduction	47
5.2. Differential Item Functioning Model	48
5.3. Basic Boosting Procedures	50
5.4. Boosting in Differential Item Functioning	53
5.4.1. The DIF Model as a Generalized Linear Model	53
5.4.2. The DIFboost Algorithm	54
5.4.3. Illustrating Example	57
5.4.4. Stability Selection	58

5.4.5. Identifiability	60
5.5. Simulation Study	61
5.5.1. Comparison to Established Methods	61
5.5.2. Simulations with Many Covariates	66
5.6. DIF in the Intelligence-Structure-Test 2000 R	68
5.7. Concluding Remarks	72
6. The Bradley–Terry Model	73
6.1. The Basic Bradley–Terry Model	73
6.2. Extensions of the Bradley–Terry Model	74
7. Modelling Heterogeneity in Paired Comparison Data	77
7.1. Introduction	77
7.2. Bradley-Terry Models with Covariates	78
7.2.1. The Basic Model	78
7.2.2. Bradley-Terry Models with Ordered Response	79
7.2.3. Heterogeneity in the Bradley-Terry Model	80
7.3. Penalized Estimation	81
7.3.1. Embedding into Generalized Linear Models	82
7.3.2. Selection by Penalization	83
7.3.3. Implementation	85
7.3.4. Choice of Penalty Parameter	86
7.3.5. Confidence Intervals	86
7.4. Application to Pre-Election Data from Germany	87
7.4.1. Data	87
7.4.2. Results	88
7.4.3. Inclusion of Twofold Interactions	94
7.5. Concluding Remarks	95
8. Extended Ordered Paired Comparison Models with Application to Football Data	97
8.1. Introduction	97
8.2. German Bundesliga	98
8.3. Ordered Paired Comparison Model with Home Advantage	99
8.3.1. The Basic Binary Bradley-Terry Model	99
8.3.2. Ordinal Models Including the Advantage in Playing at Home	100
8.3.3. Fitting the Model	102
8.3.4. Football Data	102
8.4. Identification of Clusters	105
8.4.1. Clustering of Teams	106
8.4.2. Clustering of Teams and Home Effects	107
8.5. Accounting for Explanatory Variables	110
8.5.1. A Model with Team-Specific Explanatory Variables	110
8.5.2. Evaluation of Penalized Estimation	113

8.6. Application to the Data from Bundesliga Season 2013/2014	114
8.6.1. Ranks and Abilities	114
8.6.2. Team-specific Home Effects	115
8.6.3. Identification of Clusters	116
8.6.4. Clustering of Teams and Home Effects	117
8.6.5. Accounting for Explanatory Variables	118
8.7. Concluding Remarks	119
9. Prediction of Soccer Tournaments Based on Regularized Poisson Regression	121
9.1. Introduction	121
9.2. Model and Estimation	124
9.3. Application	127
9.3.1. Data	127
9.3.2. Estimation Results	131
9.3.3. Goodness-of-fit	134
9.3.4. Prediction Power	138
9.3.5. Probabilities for FIFA World Cup 2014 Winner	139
9.3.6. Most Probable Tournament Outcome	144
9.4. Concluding Remarks	146
10. Conclusion and Outlook	147
Appendices	155
A. Visualization of Categorical Response Models	157
A.1. Introduction	157
A.2. The Multinomial Logit Model	158
A.3. Traditional Methods of Visualization: Probability Plots	160
A.4. Glyphs for the Visualization of Parameters	163
A.4.1. Star Plots for Parameters	164
A.4.2. Extensions and Alternatives	166
A.4.3. Alternative Displays	168
A.4.4. Further Examples	170
A.5. Ordinal Response Models	173
A.6. Concluding Remarks	177
B. Identifiability of the DIF Model	179
C. Additional Results for the WC1994 Data in Chapter 9	181
References	187

1. Introduction

Commonly, item response models and paired comparison models are treated as different model classes, suited for different data situations. However, there is a great similarity between item response data and paired comparisons and, accordingly, between the respective modeling approaches. Item response data appear when test persons face a certain number of items which are designed to measure a specific latent trait of the test persons. Such latent traits can, for example, be certain skills (e.g. intelligence) of the test persons or attitudes towards a specific issue (e.g. xenophobia). In the simplest case only two outcomes are possible, for example right or wrong answers or approving or disapproving of a statement.

Paired comparison data occur if two objects or items compete in a certain way. The most frequent occurrence of paired comparisons is when two objects are presented and raters have to declare a preference for one or the other object. But also in other situations paired comparisons appear, as, for example, in sport competitions between two players or teams. Again, in the simple case only two outcomes are possible, namely the win/preference of one object over the other.

Both in item response data and in paired comparisons, the outcome refers to the result of a specific competition between two actors. Therefore, item response data can be seen as a special type of paired comparison data. Tutz (1989) distinguishes between homogeneous and heterogeneous paired comparisons. In this sense, item response data are heterogeneous paired comparisons as the matched pairs are pairs of one item and one respondent. In contrast, homogeneous paired comparisons treat matched pairs of two objects or items.

The basic and most popular models for these data are the Rasch model (RM) for item response data and the Bradley-Terry or Bradley-Terry-Luce model (BTL) for paired comparison data. The Rasch model (Rasch, 1960) assumes that the probability that a person solves an item is determined by the difference between one latent parameter representing the person and one latent parameter representing the item. Let the random variable Y_{pi} represent the response where $Y_{pi} = 1$ if person p solves item i and $Y_{pi} = 0$ otherwise. With the Rasch model the probability that person p solves item i is modeled by

$$P(Y_{pi} = 1) = \frac{\exp(\theta_p - \beta_i)}{1 + \exp(\theta_p - \beta_i)} \quad p = 1, \dots, P, i = 1, \dots, I$$

where θ_p is the person parameter and β_i is the item parameter. In contrast, the Bradley-Terry model (Bradley and Terry, 1952) for a competition between two objects a_r and a_s models the probability that a_r beats a_s by

$$P(Y_{(rs)} = 1) = \frac{\exp(\gamma_r - \gamma_s)}{1 + \exp(\gamma_r - \gamma_s)}.$$

The parameters $\gamma_r, r = 1, \dots, m$, are the trait parameters of the objects $\{a_1, \dots, a_m\}$. The random variable $Y_{(rs)}$ denotes the response where $Y_{(rs)} = 1$ if object a_r is preferred over a_s and $Y_{(rs)} = 0$ otherwise.

Comparing these two basic models, "the direct relationship between the RM and the BTL is obvious" (Fischer and Molenaar, 1995). Both models are logit models, their linear predictors represent the difference between the latent traits of both actors. The main difference is, that the two actors are one item and one person for the Rasch model but two items for the Bradley-Terry model. In this thesis, both models for homogeneous and heterogeneous paired comparisons, in particular the Rasch model and the Bradley-Terry model, will be extended in various ways. The proposed extensions are supposed to allow for more flexibility in the modelling of item response and paired comparison data and for the inclusion of more information than in classical modelling approaches. A main focus will be on the inclusion of covariates.

All proposed methods will use regularization techniques for estimation. The main goal of regularization is to prevent overfitting and to allow for unique solutions in ill-posed problems, see Hastie et al. (2009) for an introduction into a broad variety of regularization methods. In this thesis, two different regularization techniques will be used, namely penalization and boosting. In penalization methods for regression models, the regular log-likelihood is maximized with respect to a certain side constraint. The resulting penalized likelihood

$$l_p(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \lambda J(\boldsymbol{\beta})$$

for a model with a general parameter vector $\boldsymbol{\beta}$ consists of the regular log-likelihood $l(\boldsymbol{\beta})$ and a penalty term $J(\boldsymbol{\beta})$ in combination with a tuning parameter λ . Famous examples for penalization methods are the ridge regression (Hoerl and Kennard, 1970) or lasso regression (Tibshirani, 1996). While ridge restricts the L_2 norm of the parameter vector $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)$ using the penalty term

$$J(\boldsymbol{\beta}) = \sum_{i=1}^p \beta_i^2,$$

lasso restricts the L_1 norm of the parameter vector with the penalty term

$$J(\boldsymbol{\beta}) = \sum_{i=1}^p |\beta_i|.$$

A main feature of penalization methods is shrinkage. The estimated coefficients are shrunk toward zero leading to a decreased variance of the estimates. In total, although the shrinkage effect goes along with biased estimates the decreased variance can lead to a decreased mean square error. Some penalization methods as, for example, the lasso also provide a dimension reduction in the covariate space. In the case of lasso, this means that lasso is able to provide parameter estimates equal to zero. Therefore, lasso allows for automatic parameter selection. In recent years, several penalty terms suited for different regression models and different data structures were developed.

Boosting evolved within the machine learning community rather than in the statistical modelling community. First approaches were proposed by Freund et al. (1996) and Tukey (1977). In the context of regression models, boosting was developed by Friedman et al. (2000) and extended, for example, by Bühlmann and Yu (2003) and Bühlmann and Hothorn (2007a). The main feature of boosting is the principle that many weak learners are combined into one joint and (hopefully) strong learner. In regression models, boosting combines many weak learners into a joint model. The main goal is to gradually improve a certain loss function, for example the L_2 loss or specific likelihood functions. In this context, a learner is considered to be a weak learner if it improves the respective loss function only by a little amount. This concept helps to avoid overfitting as the procedure is not supposed to be performed until convergence. Many boosting procedures, including the one proposed in this thesis, also allow for variable selection.

Guideline through the Thesis

This thesis consists of 10 chapters and three appendices. Chapters 2 and 3 contain general introductions into the most important topics treated in Chapters 4 and 5. Chapter 2 provides an introduction into the Rasch model, together with its most important assumptions and properties and the typical estimation methods. Chapter 3 gives a short introduction into the topic of differential item functioning. As Chapters 4 and 5 propose new methods for the detection of differential item functioning, Chapter 3 also presents some of the most popular methods for the detection of differential item functioning.

Chapter 4 proposes a new diagnostic tool for the identification of differential item functioning (DIF). In particular, an explicit model for differential item functioning is proposed that includes a set of variables. In contrast to most classical approaches to detect DIF, which

only allow to consider few (mostly two) subpopulations, the proposed model can handle both continuous and categorical covariates. The ability to include a set of covariates entails that the model contains a large number of parameters. Penalized maximum likelihood estimators are used to solve the estimation problem and to identify the items that induce DIF. It is shown that the method is able to detect items with DIF. Simulations and two applications demonstrate the applicability of the method.

Chapter 5 continues the idea from Chapter 4 to identify differential item functioning using several covariates at the same time and proposes a boosting algorithm instead of the penalized likelihood approach. The covariates can be both continuous and (multi-)categorical, and also interactions between covariates can be considered. The method works for the general parametric model for DIF in Rasch models proposed in Chapter 4. Since the boosting algorithm selects variables automatically, it is able to detect the items which induce DIF. It is demonstrated that boosting competes well with traditional methods in the case of subgroups. Furthermore, it outperforms the method proposed in Chapter 4 in the case of metric covariates. The method is illustrated by an extensive simulation study and an application to real data.

While Chapters 2-5 treat some basics and some new proposals in the context of item response data and the inclusion of covariates, the following chapters consider methods suited for paired comparison data. First, Chapter 6 introduces the basic Bradley-Terry model together with the most important existing extensions of the model.

In traditional paired comparison models heterogeneity in the population is simply ignored and it is assumed that all persons have the same preference structure. In Chapter 7, a new method to model heterogeneity in paired comparison data is proposed. The preference of an item over another item is explicitly modelled as depending on attributes of the subjects. Therefore, the model allows for heterogeneity between subjects as the preference for an item can vary across subjects depending on subject-specific covariates. Since by construction the model contains a large number of parameters we propose to use penalized estimation procedures to obtain estimates of the parameters. The used regularized estimation approach penalizes the differences between the parameters corresponding to single covariates. It enforces variable selection and allows to find clusters of items with respect to covariates. We consider simple binary but also ordinal paired comparisons models. The method is applied to data from a pre-election study from Germany.

In Chapter 8, a general paired comparison model for the evaluation of sport competitions is proposed. It efficiently uses the available information by allowing for ordered response categories and team-specific home advantage effects. Penalized estimation techniques are used to identify clusters of teams that share the same ability. The model is extended to include team-specific explanatory variables. Therefore, in contrast to Chapter 7, object-specific covariates are considered instead of subject-specific covariates. It is shown that regularization

techniques allow to identify the contribution of team-specific covariates to the success of teams. The usefulness of the method is demonstrated by investigating the performance and its dependence on the budget for football teams of the German Bundesliga.

In Chapter 9 an approach for the analysis and prediction of international soccer match results is proposed. In contrast to Chapter 8, the result of one match is not modeled as an ordered response. Instead, the number of scored goals is modeled directly using a Poisson distribution. To account for the paired comparison structure of the data, the linear predictor consists of differences between the covariates of both competing teams. Therefore, similar as in Chapter 8 object-specific covariates are included in the model. Lasso approaches are used to achieve variable selection and shrinkage. Based on preceding FIFA World Cups, two models for the prediction of the FIFA World Cup 2014 are fitted and investigated. Based on the model estimates, the FIFA World Cup 2014 is simulated repeatedly and winning probabilities are obtained for all teams. Both models favor the actual FIFA World Champion Germany.

In Chapters 4 and 5 the concept of effect stars is used to visualize parameter estimates for DIF items, in chapter 7 effect stars are used to visualize estimates from the proposed method BTLasso. Originally, effect stars were proposed to visualize parameter estimates in categorical response models, in particular for multinomial and ordinal logit models. Therefore, in Appendix A the original concept of effect stars in the context of multinomial logit models is introduced. The multinomial logit model is the most widely used model for nominal multi-category responses. One problem with the model is that many parameters are involved, another that interpretation of parameters is much harder than for linear models because the model is non-linear. Both problems can profit from graphical representations. Effect stars visualize the effect strengths by star plots, where one star collects all the parameters connected to one term in the linear predictor. In contrast to conventional star plots, which are used to represent data, the plots represent parameters and are considered as parameter glyphs. The method is extended to ordinal models and illustrated by several data sets.

In order to keep the single chapters self-contained, every chapter contains separate introductions to the relevant topics and a separate conclusion. Therefore, every chapter can also be read separately but some topics will repeat themselves.

Contributing Manuscripts

Parts of this thesis were published as articles in peer reviewed journals, other parts were published in proceedings of scientific conferences or as technical reports at the Department of Statistics of the Ludwig-Maximilians-Universität München. In the following, chapter by

chapter all contributing manuscripts are listed together with a declaration of the personal contributions of the respective authors:

Chapter 4: Tutz and Schauburger (2015b). A Penalty Approach to Differential Item Functioning in Rasch Models. *Psychometrika* 80(1), 21 – 43

The project was initiated by Gerhard Tutz and further developed jointly by Gerhard Tutz and Gunther Schauburger. Gunther Schauburger implemented the method and performed the simulations and the real data analyses. Gunther Schauburger developed the corresponding R-package `DIFlasso`. The manuscript was written close collaboration of both authors. The original manuscript is extended by Section 4.6, which discusses concepts of variable selection within the proposed method. Apart from this section and some minor modifications Chapter 4 and Tutz and Schauburger (2015b) match. The technical report 134 (Tutz and Schauburger, 2012a) and the conference paper from the IWSM 2013 (Schauburger and Tutz, 2013) contain preliminary work on the project.

Chapter 5: Schauburger and Tutz (2015b). Detection of Differential Item Functioning in Rasch Models by Boosting Techniques. *British Journal of Mathematical and Statistical Psychology*, published online

The project was initiated jointly by Gerhard Tutz and Gunther Schauburger. Main author of the manuscript was Gunther Schauburger in close collaboration with Gerhard Tutz. Gunther Schauburger was responsible for the implementation of the method, of the simulation studies and the application to real data. Furthermore, Gunther Schauburger developed the corresponding R-package `DIFboost`. Apart from minor modifications Chapter 5 and Schauburger and Tutz (2015b) match. The conference paper from the IWSM 2014 (Schauburger and Tutz, 2014) contains preliminary work on the project.

Chapter 7: Schauburger and Tutz (2015c). Modelling Heterogeneity in Paired Comparison Data – an L_1 Penalty Approach with an Application to Party Preference Data. *Department of Statistics, LMU Munich*, Technical Report 183

The project was initiated and realized in close collaboration. Gunther Schauburger as the first author mainly wrote most of the manuscript and performed the presented analyses. He was also responsible for the implementation of the method and the corresponding R-package `BTLlasso`. The original manuscript is extended by Subsection 7.4.3 which discusses the inclusion of twofold interactions in the application and by a paragraph applying the concept of effect stars to the estimates of the proposed method. Apart from these extensions and minor modifications Chapter 7 and Schauburger and Tutz (2015c) match. The conference

paper from the IWSM 2015 (Schauberger and Tutz, 2015a) contains preliminary work on the project.

Chapter 8: Tutz and Schaubberger (2015a). Extended Ordered Paired Comparison Models with Application to Football Data from German Bundesliga. *Advances in Statistical Analysis*, 99(2), 209 – 227

The manuscript was a joint project of Gerhard Tutz and Gunther Schaubberger. Both authors contributed to the manuscript. The data collection and all implementations were done by Gunther Schaubberger. The original manuscript is extended by Section 8.6 where the analyses from the previous sections are applied to the data from another Bundesliga season. Apart from this section and minor modifications Chapter 8 and Tutz and Schaubberger (2015a) match. The technical report 151 (Tutz and Schaubberger, 2014) contains preliminary work on the project.

Chapter 9: Groll, Schaubberger, and Tutz (2015). Prediction of Major International Soccer Tournaments Based on Team-Specific Regularized Poisson Regression: An Application to the FIFA World Cup 2014. *Journal of Quantitative Analysis in Sports* 11(2), 97 – 115

Andreas Groll and Gunther Schaubberger initiated and conducted the project in close collaboration. In particular, they were equally responsible for the data collection, the implementation of the methods and the manuscript. Gerhard Tutz supervised the methodological part of the manuscript and helped to improve the manuscript by extensive discussions. Apart from minor modifications Chapter 9 and Groll et al. (2015) match. The technical report 166 (Groll et al., 2014) contains preliminary work on the project.

Appendix A: Tutz and Schaubberger (2013): Visualization of Categorical Response Models: From Data Glyphs to Parameter Glyphs. *Journal of Computational and Graphical Statistics*, 22(1), 156 – 177

The manuscript was mainly drafted by Gerhard Tutz with contributions of Gunther Schaubberger. Gunther Schaubberger was responsible for the implementation including the corresponding R package *EffectStars* (Schauberger, 2014b) and for all visualizations in the manuscript. He was strongly involved in all parts of the final manuscript. Apart from minor modifications Appendix A and Tutz and Schaubberger (2013) match. The technical report 117 (Tutz and Schaubberger, 2012a) and the conference paper from the IWSM 2012 (Schauberger and Tutz, 2012) contain preliminary work on the project.

Software

Most computations in this thesis were done with the statistical program **R** (R Core Team, 2015), parts were implemented in **C++** but are integrated in **R**. For most of the methods proposed in this thesis add-on packages for **R** were developed which can be downloaded from the Comprehensive R Archive Network (CRAN). In particular, the following **R**-packages were developed:

DIFlasso provides the method DIFlasso proposed in Chapter 4 (Schauberger, 2014a).

DIFboost provides the method DIFboost proposed in Chapter 5 (Schauberger, 2015b).

BTLLasso provides the method BTLLasso proposed in Chapter 7 (Schauberger, 2015a).

The fitting algorithm of **BTLLasso** is implemented in **C++** code which is integrated into **R** using the packages **Rcpp** (Eddelbuettel, 2013) and **RcppArmadillo** (Eddelbuettel and Sanderson, 2014).

EffectStars provides the concept of effect stars proposed in Appendix A (Schauberger, 2014b).

2. The Rasch Model

In the following, the basic Rasch model (Rasch, 1960) will be explained in more detail. The Rasch model is considered to be a starting point of the item response theory (IRT) which over the last decades replaced the classical test theory (CTT) as the most popular method in the analysis of tests or questionnaires in general. The main difference between the CTT and the IRT is that the IRT models a probabilistic distribution of the correct response probability. The most general IRT model is the so called 3PL model (Birnbaum, 1968). It models the probability of a specified response depending on item parameters and a person parameter. Typically, such a specified response will simply be the (either correct or wrong) answer on a test question. If person p , $p = 1, \dots, P$, tries to solve item i , $i = 1, \dots, I$, the response is denoted as

$$Y_{pi} = \begin{cases} 1 & \text{person } p \text{ solves item } i \\ 0 & \text{otherwise} \end{cases}$$

Accordingly, the 3PL model is denoted by

$$P(Y_{pi} = 1) = c_i + (1 - c_i) \frac{\exp(a_i(\theta_p - \beta_i))}{1 + \exp(a_i(\theta_p - \beta_i))}.$$

Here, θ_p represents the person ability and β_i represents the item difficulty. The parameters c_i and a_i represent the guessing parameter and the discrimination parameter of item i . The model is called 3PL model as one item i is characterized by three item parameters, a_i, β_i, c_i . From the 3PL model, the 2PL model and the 1PL model can be obtained as special cases. In the 2PL model, it is assumed that no guessing is possible and the restriction $c_i = 0$, $i = 1, \dots, I$ is applied. In the 1PL model (in the following referred to as the Rasch model), additionally equal discrimination parameters are assumed by restricting $a_i = 1$, $i = 1, \dots, I$.

In the analysis of item response data, the Rasch model is the most popular choice. If person p , $p = 1, \dots, P$, tries to solve item i , $i = 1, \dots, I$, this is specified by the Rasch model by

$$P(Y_{pi} = 1) = \frac{\exp(\theta_p - \beta_i)}{1 + \exp(\theta_p - \beta_i)}$$

where θ_p represents the latent person ability and β_i represents the latent item difficulty. For identifiability, a restriction on the parameters is needed. Frequently, either a person parameter or an item parameter is set zero. Basically, the Rasch model simply represents a binomial logit model and can, therefore, easily be embedded into the framework of generalized linear models (GLMs) (McCullagh and Nelder, 1989). The Rasch model makes the person abilities and the item difficulties comparable. For example, if the ability of person p equals the difficulty of item i (i.e. $\theta_p = \beta_i$), the Rasch model will predict a probability of 0.5 that person p will solve item i .

2.1. Assumptions and Properties of the Rasch Model

The Rasch model is accompanied by four main assumptions, namely monotonicity, unidimensionality, conditional independence and sufficiency, compare Hatzinger (1989) and Kelderman (1984).

Monotonicity The solving probability $P(Y_{pi} = 1|\theta_p, \beta_i)$ is strictly monotone increasing for $\theta_p \in \mathbb{R}$. Furthermore, $P(Y_{pi} = 1|\theta_p, \beta_i) \rightarrow 0$ for $\theta_p \rightarrow -\infty$ and $P(Y_{pi} = 1|\theta_p, \beta_i) \rightarrow 1$ for $\theta_p \rightarrow \infty$ holds. Therefore, with increasing ability, the probability to solve an item increases.

Unidimensionality Given the item difficulty, the probability to solve an item solely depends on the true value of the respective person on the latent trait. That means that $P(Y_{pi} = 1|\theta_p, \beta_i, \phi) = P(Y_{pi} = 1|\theta_p, \beta_i)$ holds for any additional variable ϕ . Given the ability parameter and the item difficulty, the solving probability does not depend on any other variables ϕ .

Conditional independence Given the latent trait, the items have to be stochastically independent. Therefore, for equally able persons the solving probabilities for different items are independent. Solving one item does not increase or decrease the probability to solve another item. Conditional independence is also widely known as local independence.

Sufficiency The total score of a person $S_p = \sum_i Y_{pi}$ contains the entire information for the ability of the person. The score is a sufficient statistic for the person parameter θ_p , persons with the same score have the same ability. Accordingly, also the number of persons that solved an item i , namely $R_i = \sum_p Y_{pi}$, is a sufficient statistic for the item difficulty.

In the Rasch model (as in all IRT models), items can be visualized using so-called item characteristic curves (ICCs). An ICC shows the probability of a correct response on the respective item depending on the person parameter θ_p . Figure 2.1 exemplarily shows the

ICCs for three items with different item difficulties. The main feature of ICCs in Rasch

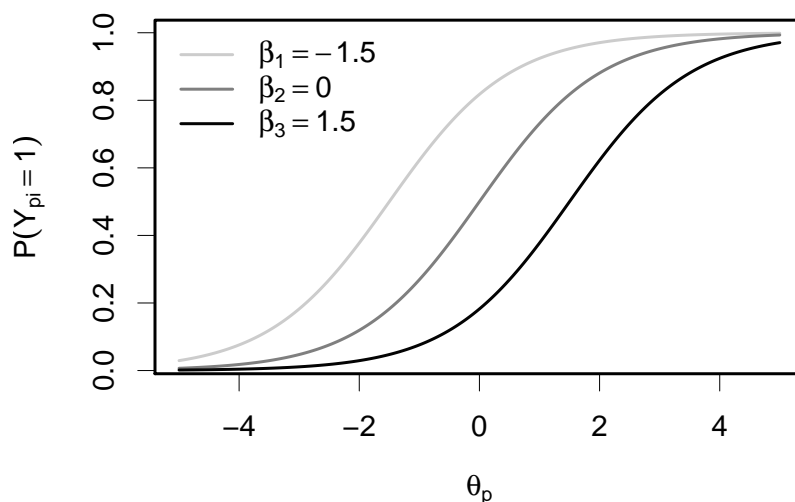


Figure 2.1.: Exemplary item characteristic curves for three items in a Rasch model

models is that they all share the same form (they have the same slope) and are only shifted vertically depending on the respective item difficulty.

2.2. Estimation Approaches for the Rasch Model

To estimate the Rasch model, three different maximum likelihood approaches exist: Joint maximum likelihood (JML), conditional maximum likelihood (CML) and marginal maximum likelihood (MML). JML simultaneously provides estimates both for the person parameters and the item parameters. CML and MML only provide item parameters, person parameters have to be estimated separately.

Joint Maximum Likelihood Estimation

The joint maximum likelihood estimation of Rasch models is the easiest and most intuitive estimation method. If an appropriate design matrix is built, it can easily be performed using standard software for GLMs. Using the restriction $\theta_P = 0$, the design matrix can be seen from

$$\log \left(\frac{P(Y_{pi} = 1)}{P(Y_{pi} = 0)} \right) = \theta_p - \beta_i = \mathbf{1}_{P(p)}^\top \boldsymbol{\theta} - \mathbf{1}_{I(i)}^\top \boldsymbol{\beta} = \mathbf{x}_{pi}^\top \boldsymbol{\delta},$$

where $\mathbf{1}_{P(p)}^\top = (0, \dots, 0, 1, 0, \dots, 0)$ has length $P - 1$ with 1 at position p , $\mathbf{1}_{I(i)}^\top = (0, \dots, 0, 1, 0, \dots, 0)$ has length I with 1 at position i , and the parameter vectors are $\boldsymbol{\theta} = (\theta_1, \dots, \theta_{P-1})$, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_I)$ yielding the total vector $\boldsymbol{\delta}^\top = (\boldsymbol{\theta}^\top, \boldsymbol{\beta}^\top)$. The design vector linked to person p and item i is given by $\mathbf{x}_{pi}^\top = (\mathbf{1}_{P(p)}^\top, -\mathbf{1}_{I(i)}^\top)$. Finally, the Rasch model can be estimated by combining all single design vectors \mathbf{x}_{pi} into a design matrix and by stacking all responses Y_{pi} appropriately into a response vector.

Estimation based on JML faces two main problems. First, if a person solves all or no items, its ability estimate will diverge to $\theta_p = \infty$ or $\theta_p = -\infty$, respectively. Equivalently, items that were solved by all or no persons will not have finite estimates although this case is much more unlikely as in general the number of persons clearly exceeds the number of items. After all, in both cases the respective person or item has to be removed from the design matrix. Second, the estimates for the item parameters from JML are inconsistent and biased for $P \rightarrow \infty$ and I fixed, see e.g. Andersen (1973b, 1980). Therefore, in recent years JML is decreasingly used in practice.

Conditional Maximum Likelihood Estimation

Nowadays, the conditional maximum likelihood method is the most popular choice. It is based on the property, that the sum score $S_p = \sum_i Y_{pi}$ of a person p is sufficient for the ability θ_p of person p . When conditioning on the sum scores the solving probabilities only depend on the item difficulties. Therefore, CML initially only provides estimates for the item parameters. Based on the item parameters, estimates for the person parameters can be obtained in a second step.

Let $\mathbf{y}_p = (y_{p1}, \dots, y_{pI})$ represent the response pattern of person p with the corresponding sum score $s_p = \sum_i y_{pi}$. Following Hatzinger (1989), the probability to observe the pattern \mathbf{y}_p , conditional on the respective sum score S_p , is denoted by

$$\begin{aligned} P(\mathbf{Y}_p = \mathbf{y}_p | S_p = s_p) &= \frac{P(\mathbf{Y}_p = \mathbf{y}_p)}{P(S_p = s_p)} \\ &= \frac{\exp(\theta_p s_p) \exp(-\sum_i \beta_i y_{pi}) / \prod_p (1 - \exp(\theta_p - \beta_i))}{\exp(\theta_p s_p) \gamma(s_p; \beta_1, \dots, \beta_I) / \prod_p (1 - \exp(\theta_p - \beta_i))}. \end{aligned} \quad (2.1)$$

Here, $\gamma(s_p; \beta_1, \dots, \beta_I) = \sum_{\mathbf{y}|s_p} \exp(-\sum_i \beta_i y_{pi})$ represents the elementary symmetric function and $\mathbf{y}|s_p$ represents all possible response patterns with a sum score s_p . It can be seen

that all terms depending on θ_p can be eliminated from (2.1). Combining all possible sum scores $t = 0, \dots, I$, the conditional likelihood can finally be denoted by

$$L_c = \frac{\exp(-\sum_i \beta_i r_i)}{\prod_t \gamma(t; \beta_1, \dots, \beta_I)^{n_t}},$$

where $r_i = \sum_p y_{pi}$ denotes the number of persons that solved item i and n_t is the number of subjects with $s_p = t$. Maximizing the conditional likelihood provides consistent estimates for the item parameters when $P \rightarrow \infty$. Afterwards, the person parameters can be estimated assuming the item parameters to be fixed. The conditional maximum likelihood shares the problem of the joint maximum likelihood that for items solved by all or no persons and for persons that solved all or no items, no finite estimates can be found.

Marginal Maximum Likelihood Estimation

Similar to the conditional maximum likelihood approach, the marginal likelihood approach uses the trick to estimate the item parameters separately by eliminating the person parameters from the likelihood. In the case of the marginal likelihood, this is done by assuming a certain distribution for the person parameters. Typically, the person parameters are assumed to be normally distributed. With a given distribution, the person parameters can be integrated out from the likelihood.

The person parameters are assumed to be a random sample of the distribution $G(\theta)$. Then the probability to observe the pattern \mathbf{y}_p can be denoted by

$$P(\mathbf{Y}_p = \mathbf{y}_p) = \int_{-\infty}^{\infty} P(\mathbf{y}_p | \theta_p) dG(\theta_p).$$

Using the parameters of the Rasch model, this can be denoted by

$$P(\mathbf{Y}_p = \mathbf{y}_p) = \exp(\beta_i r_i) \int_{-\infty}^{\infty} \frac{\exp(\theta_p s_p)}{\prod_{i=1}^I (1 - \exp(\theta_p - \beta_i))} dG(\theta_p).$$

Finally, the marginal likelihood is defined as product over all persons of the probability above. Then, the likelihood is a function depending on the item parameters and the distribution $G(\theta)$ and can be maximized with regard to the respective parameters. Due to the distributional assumption, using the marginal likelihood the estimates for the persons with perfect scores or scores of zero are finite.

3. Differential Item Functioning

Psychological or educational tests are typically used to investigate a latent trait of a person like the intelligence or other specific skills. For this purpose, appropriate items are needed to provide a valid measurement of the respective trait. Items are considered to be unfair if, for a specific item, two persons with the same underlying latent trait have different probabilities to answer the item correctly. Then, the item functions differently for two persons with the same value of the latent trait. Therefore, this phenomenon is called differential item functioning (DIF). In former publications, DIF was also denominated by the terms measurement bias or item bias, see, e.g., Lord (1980), Swaminathan and Rogers (1990) or Millsap and Everson (1993). Nowadays, the more neutral term of differential item functioning has widely prevailed.

Over the past decades, a vast amount of methods has been proposed to detect DIF. For an overview of the most popular methods see, e.g., Holland and Wainer (2012), Millsap and Everson (1993) or, more up to date, Magis et al. (2010). Typically, DIF is investigated by testing if special (known) characteristics of the participants like gender or ethnicity alter the probability to score on an item. Alternatively, also (unknown) latent classes could be assumed to describe DIF as proposed by Rost (1990). Here one assumes, that a model holds for all persons within a latent class but models for different classes differ. Since it is unclear what the latent classes represent, interpretation is rather hard and much less intuitive than for DIF between known groups. Therefore, latent class models have not become an established tool in DIF detection.

DIF can be divided into uniform and nonuniform DIF. Uniform DIF means that the difference between the solving probabilities for an item is constant along the person abilities for two equally able persons. In nonuniform DIF, the magnitude of the DIF effect depends on the respective person ability. Figure 3.1 exemplarily shows the item characteristic curves for items with uniform (left) and nonuniform (right) DIF between two subgroups of the population. It can be seen, that for nonuniform DIF the item characteristic curves can also be crossing. While the item is easier for group 1 than for group 2 on a low ability level it is harder on a high ability level. Within the context of IRT models, nonuniform DIF can be found in 2PL or 3PL models because only here the item characteristic curves can have different slopes. In case of the Rasch model introduced in Chapter 2, only uniform DIF is possible as all discrimination parameters are assumed to be fixed $a_i = 1$, $i = 1, \dots, I$.

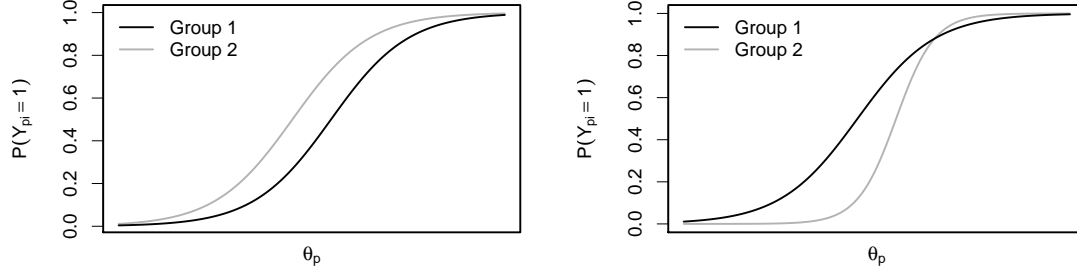


Figure 3.1.: Exemplary item characteristic curves for an item with uniform (left) and nonuniform (right) DIF between two subgroups

3.1. Popular Methods for DIF Detection

Traditional DIF methods are focused on the detection of DIF between two groups (reference and focal group) of participants, typically males and females. In the following, the most popular methods for the case of two group comparison are presented.

For DIF detection, the Mantel-Haenszel (MH) method has become a popular choice. It was proposed by Holland and Thayer (1988) and is named by the Mantel-Haenszel statistic proposed by Mantel and Haenszel (1959). The method can be described as an item-wise contingency table method and is not based on an underlying IRT model. It can not detect nonuniform DIF but only uniform DIF. The participants of the test are matched by their total test score. For each distinct test score, a 2×2 table collects the number of correct and incorrect answers for the reference and the focal group. Finally, a χ^2 -statistic can be calculated across all contingency tables. A χ^2 -test with one degree of freedom is performed.

Swaminathan and Rogers (1990) proposed the method of logistic regression which is based on the item-specific model

$$\log \left(\frac{P(Y_{pi} = 1)}{P(Y_{pi} = 0)} \right) = \beta_0 + \beta_1 S_p + \beta_2 G_p + \beta_3 S_p G_p,$$

where G_p represents the group membership (0: reference group, 1: focal group) and S_p represents the total test score of person p . The parameters β_2 and β_3 can then be tested by Wald or Likelihood-Ratio-Tests. Testing β_2 represents a test on uniform DIF while a test of β_3 represents a test on nonuniform DIF. If only uniform DIF is investigated, the interaction $S_p G_p$ can be omitted.

In contrast to both aforementioned approaches, the method of Lord (Lord, 1980) is explicitly based on IRT models. In the case of two subgroups, separate IRT models are fitted for the reference group and the focal group. Then, the differences of the item parameters are tested using a χ^2 -test. If the difference between the item parameters corresponding to the same item differs significantly from zero, the item is diagnosed as DIF item. In general, the corresponding test statistic for item i is denoted by $Q_i = (\alpha_{iR} - \alpha_{iF})^T (\Sigma_{iR} - \Sigma_{iF})^{-1} (\alpha_{iR} - \alpha_{iF})$ where $\alpha_{iR} = (a_{iR}, \beta_{iR}, c_{iR})$ and $\alpha_{iF} = (a_{iF}, \beta_{iF}, c_{iF})$ collect all item parameters from the models for the reference group and the focal group, respectively. Σ_{iR} and Σ_{iF} represent the variance-covariance matrices of the respective estimates. In the case of the Rasch model (1PL model), the test statistic is reduced to $Q_i = (\beta_{iR} - \beta_{iF})^2 / (\hat{\sigma}_{iR}^2 + \hat{\sigma}_{iF}^2)$ with $\hat{\sigma}_{iR}^2$ and $\hat{\sigma}_{iF}^2$ representing the estimated variances of the difficulty estimates β_{iR} and β_{iF} . In that case, the detection of DIF is restricted to uniform DIF. For the method of Lord, asymptotic normality of the item parameters is assumed.

All three aforementioned methods have been extended to the case of multiple groups instead of one reference and one focal group. Somes (1986) and Penfield (2001) extended the MH approach, Magis et al. (2011) extended the logistic regression method and Kim et al. (1995) extended the method of Lord. All methods (together with their extension to multiple groups) are implemented in the R-package `difR` (Magis et al., 2010, 2013).

3.2. Problems and Limitations

The presented methods of DIF detection share two main problems (Millsap and Everson, 1993). First, the tests are designed for only one covariate. In the case of multiple covariates several tests have to be performed at the same time and the problem of multiple testing arises. Therefore, correction strategies like the Bonferroni correction have to be applied (e.g., Penfield, 2001). Second, the participants from different groups are matched by their total test scores, assuming that the total test score is an appropriate measurement of the latent traits of the participants. Therefore, one assumes that all other items (except for the studied item) do not show DIF and are treated as so-called anchor items. Generally, all DIF methods depend on anchor items. They are assumed to be DIF-free and serve as references for the item under investigation. A contaminated set of anchor items (i.e. the set contains DIF items) will lead to an increased error rate in DIF identification. For an overview of different anchor strategies, see Kopf et al. (2015). Obviously, for a method of DIF detection the assumption that all other items are DIF-free is rather paradoxical. Especially if several items show DIF in favor of the same group, the test scores can be considerably biased. Then, the test score is no longer a fair measurement and the quality of the DIF method will suffer. Therefore, an unbiased measure of the latent trait is needed. For this purpose, so-called purification methods have been proposed, for example

by Holland and Thayer (1988), Candell and Drasgow (1988) or Clauser et al. (1993). The goal of these procedures is to iteratively detect DIF items and to exclude these items from the calculation of the test score. Purification procedures are supposed to lead to an uncontaminated measurement of the latent trait and to an improved selection performance of the respective DIF detection method. In several publications, e.g. Wang and Su (2004) or Fidalgo et al. (2000), purification methods turned out to provide a significant improvement of the respective methods.

4. A Penalty Approach to Differential Item Functioning in Rasch Models

4.1. Introduction

Differential item functioning (DIF) is the well known phenomenon that the probability of a correct response among equally able persons differs in subgroups. For example, the difficulty of an item may depend on the membership to a racial, ethnic or gender subgroup. Then the performance of a group can be lower because these items are related to specific knowledge that is less present in this group. The effect is measurement bias and possibly discrimination, see, for example, Millsap and Everson (1993), Zumbo (1999). Various forms of differential item functioning have been considered in the literature, see, for example, Holland and Wainer (2012), Osterlind and Everson (2009); Rogers (2005); Osterlind and Everson (2009). Magis et al. (2010) give an instructive overview of the existing DIF detection methods.

In this chapter we will investigate DIF in item response models, focusing on the Rasch model. In item response models DIF is considered to be uniform, that is the probability of correctly answering is uniformly greater for specific subgroups. Test statistics for the identification of uniform DIF have been proposed, among others, by Thissen et al. (1993), Lord (1980), Holland and Thayer (1988), Kim et al. (1995) and Raju (1988). More recently, DIF has been embedded into the framework of mixed models (Van den Noortgate and De Boeck, 2005) and Bayesian approaches have been developed (Soares et al., 2009). Also the test concepts developed in Merkle and Zeileis (2013) could be helpful to investigate dependence of responses on subgroups.

A severe limitation of existing approaches is that they are typically limited to the consideration of few subgroups. Most often, just two subgroups have been considered with one group being fixed as the reference group. The objective of the present chapter is to provide tools

This chapter is a modified version of Tutz and Schauburger (2015b), previous work on the issue can be found in the technical report 134 (Tutz and Schauburger, 2012a) and the conference paper Schauburger and Tutz (2013). See Chapter 1 for more information on the personal contributions of all authors and textual matches.

that allow for several groups but also for continuous variables like age to induce differential item functioning. We propose a model that lets the item difficulties to be modified by a set of variables that can potentially cause DIF. The model necessarily contains a large number of parameters which raises severe estimation problems. But estimation problems can be solved by regularized estimation procedures. Although alternative strategies could be used we focus on regularization by penalization, using penalized maximum likelihood (ML) estimates. The procedure allows to identify the items that suffer from DIF and investigate which variables are responsible.

More recently, Strobl et al. (2015) proposed a new approach that is also able to handle several groups and continuous variables but uses quite different estimation procedures. The proposed method will be compared to this alternative approach.

Chapter 4 is organized as follows. In Section 4.2 we present the model, in Section 4.3 we show how the model can be estimated. Then we illustrate the fitting of the model by use of simulation studies and real data examples.

4.2. Differential Item Functioning Model

We will first consider the binary Rasch model and then introduce a general parametric model for differential item functioning.

4.2.1. The Binary Rasch Model

The most widespread item response model is the binary Rasch model (Rasch, 1960). It assumes that the probability that a participant in a test scores on an item is determined by the difference between two latent parameters, one representing the person and one representing the item. In assessment tests the person parameter refers to the ability of the person and the item parameter to the difficulty of the item. More generally the person parameter refers to the latent trait the test is supposed to measure. With $Y_{pi} \in \{0, 1\}$ the probability that person p solves item i is given by

$$P(Y_{pi} = 1) = \frac{\exp(\theta_p - \beta_i)}{1 + \exp(\theta_p - \beta_i)} \quad p = 1, \dots, P, \quad i = 1, \dots, I$$

where θ_p is the person parameter (ability) and β_i is the item parameter (difficulty). A more convenient form of the model is

$$\log \left(\frac{P(Y_{pi} = 1)}{P(Y_{pi} = 0)} \right) = \theta_p - \beta_i, \quad (4.1)$$

where the left hand side represents the so-called logits, $\text{Logit}(P(Y_{pi} = 1)) = \log(P(Y_{pi} = 1)/P(Y_{pi} = 0))$. It should be noted that the parameters are not identifiable. Therefore, one has to fix one of the parameters. We choose $\theta_P = 0$, which yields a simple representation of the models to be considered later.

Under the usual assumption of conditional independence given the latent traits the maximum likelihood (ML) estimates can be obtained within the framework of generalized linear models (GLMs). GLMs for binary responses assume that the probability $\pi_{pi} = P(Y_{pi} = 1)$ is given by $g(\pi_{pi}) = \mathbf{x}_{pi}^T \boldsymbol{\delta}$, where $g(\cdot)$ is the link function and \mathbf{x}_{pi} is a design vector linked to person p and item i . The link function is directly seen from model representation (4.1). The design vector, which codes the persons and items and the parameter vector are seen from

$$\log \left(\frac{P(Y_{pi} = 1)}{P(Y_{pi} = 0)} \right) = \theta_p - \beta_i = \mathbf{1}_{P(p)}^T \boldsymbol{\theta} - \mathbf{1}_{I(i)}^T \boldsymbol{\beta},$$

where $\mathbf{1}_{P(p)}^T = (0, \dots, 0, 1, 0, \dots, 0)$ has length $P - 1$ with 1 at position p , $\mathbf{1}_{I(i)}^T = (0, \dots, 0, 1, 0, \dots, 0)$ has length I with 1 at position i , and the parameter vectors are $\boldsymbol{\theta} = (\theta_1, \dots, \theta_{P-1})$, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_I)$ yielding the total vector $\boldsymbol{\delta}^T = (\boldsymbol{\theta}^T, \boldsymbol{\beta}^T)$. The design vector linked to person p and item i is given by $\mathbf{x}_{pi}^T = (\mathbf{1}_{P(p)}^T, -\mathbf{1}_{I(i)}^T)$.

GLMs are extensively investigated in McCullagh and Nelder (1989), short introductions with the focus on categorical data are found in Agresti (2002) and Tutz (2012). The embedding of the Rasch model into the framework of generalized linear models has the advantage that software that is able to fit GLMs and extensions can be used to fit models very easily.

4.2.2. A General Differential Item Functioning Model

In a general model that allows the item parameters to depend on covariates that characterize the person we will replace the item parameter by a linear form that includes a vector of explanatory variables. Let \mathbf{x}_p be a person-specific parameter that contains, for example, gender, race, but potentially also continuous covariates like age. If β_i is replaced by $\beta_i + \mathbf{x}_p^T \boldsymbol{\gamma}_i$ with item-specific parameter $\boldsymbol{\gamma}_i$ one obtains the model

$$\log \left(\frac{P(Y_{pi} = 1)}{P(Y_{pi} = 0)} \right) = \theta_p - (\beta_i + \mathbf{x}_p^T \boldsymbol{\gamma}_i) \quad (4.2)$$

For illustration, let us consider the simple case where the explanatory variable codes a subgroup like gender, which has two possible values. Let $x_p = 1$ for males and $x_p = 0$ for females. If item i functions differently in the subgroups, one has the item parameters

$$\beta_i + \gamma_i \text{ for males and } \beta_i \text{ for females.}$$

Then γ_i represents the difference of item difficulty between males and females. If one prefers a more symmetric representation one can choose $x_p = 1$ for males and $x_p = -1$ for females obtaining

$$\beta_i + \gamma_i \text{ for males and } \beta_i - \gamma_i \text{ for females.}$$

Then γ_i represents the deviation of the sub-populations in item difficulty from the baseline difficulty β_i . Of course in an item that does not suffer from differential item functioning, one has $\gamma_i = 0$ and, therefore, items for males and females are equal.

The strength of the general model (4.2) is that also continuous covariates like age can be included. Thinking of items that are related to knowledge on computers or modern communication devices the difficulty may well vary over age. One could try to build more or less artificial age groups, or, as we do, assume linear dependence of the logits. With x_p denoting age in years the item parameter is $\beta_i + \text{age}\gamma_i$. If $\gamma_i = 0$ the item difficulty is the same for all ages.

The multi-group case is easily incorporated by using dummy-variables for the groups. Let R denote the group variable, for example, race with k categories, that is, $R \in \{1, \dots, k\}$. Then one builds a vector $(x_{R(1)}, \dots, x_{R(k-1)})$, where components are defined by $x_{R(j)} = 1$ if $R = j$ and $x_{R(j)} = 0$ otherwise. The corresponding parameter vector γ_i has $k - 1$ components $\gamma_i^T = (\gamma_{i1}, \dots, \gamma_{i,k-1})$. Then the parameters are

$$\beta_i + \gamma_{i1} \text{ in group 1, } \dots \beta_i + \gamma_{i,k-1}, \text{ in group } k - 1 \quad \beta_i \text{ in group } k.$$

In this coding the last category, k , serves as reference category, and the parameters $\gamma_{i1}, \dots, \gamma_{i,k-1}$ represent the deviations of the subgroups with respect to the reference category.

One can also use symmetric coding where one assumes $\sum_{j=1}^k \gamma_{ij} = 0$ yielding parameters

$$\beta_i + \gamma_{i1} \text{ in group 1, } \dots \beta_i + \gamma_{i,k-1}, \text{ in group } k - 1 \quad \beta_i + \gamma_{ik} \text{ in group } k.$$

In effect one is just coding a categorical predictor in 0 – 1-coding or effect coding, see, for example, Tutz (2012).

The essential advantage of model (4.2) is that the person-specific parameter includes all the candidates that are under suspicion to induce differential item functioning. Thus one has

a vector that contains age, race, gender and all the other candidates. If one component in the vector γ_i is unequal zero the item is group-specific, the parameter shows which of the variables is responsible for the differential item functioning. The model includes not only several grouping variables but also continuous explanatory variables.

The challenge of the model is to estimate the large number of parameters and to determine which parameters have to be considered as unequal zero. The basic assumption is that most of the parameters do not depend on the group, but some can. One wants to detect these items and know which one of the explanatory variables is responsible. For the estimation one has to use regularization techniques that are discussed in Section 3.

Identifiability Issues

The general model uses the predictor $\eta_{pi} = \theta_p - \beta_i - \mathbf{x}_p^T \gamma_i$ when person p tries to solve item i . Even if one of the basic parameters is fixed, say, $\beta_I = 0$, the model can be reparameterized by use of a fixed vector \mathbf{c} in the form

$$\eta_{pi} = \theta_p - \beta_i - \mathbf{x}_p^T \gamma_i = \theta_p - \beta_i - \mathbf{x}_p^T (\gamma_i - \mathbf{c}) - \mathbf{x}_p^T \mathbf{c} = \tilde{\theta}_p - \beta_i - \mathbf{x}_p^T \tilde{\gamma}_i,$$

where $\tilde{\theta}_p = \theta_p - \mathbf{x}_p^T \mathbf{c}$ and $\tilde{\gamma}_i = \gamma_i - \mathbf{c}$. The parameter sets $\{\theta_p, \beta_i, \gamma_i\}$ and $\{\tilde{\theta}_p, \beta_i, \tilde{\gamma}_i\}$ describe the same model, the parameters are just shifted by $\mathbf{x}_p^T \mathbf{c}$ in the case of θ -parameters and \mathbf{c} in the case of γ -parameters. In other words, the model is overparameterized and parameters are not identifiable. Additional constraints are needed to make the parameters identifiable. However, the choice of the constraints determines which items are considered as DIF-inducing items. Let us consider a simple example with a binary variable \mathbf{x}_p , which codes, for example, gender. Then the parameters are identifiable if one sets one β -parameter and one γ -parameters to zero. With six items and the unconstrained parameters $(\gamma_1, \dots, \gamma_6) = (5, 5, 5, 3, 3, 3)$ the constraint $\gamma_1 = 0$ yields the identifiable parameters $(\gamma_1, \dots, \gamma_6) = (0, 0, 0, -2, -2, -2)$, whereas the constraint $\gamma_6 = 0$ yields the identifiable parameters $(\gamma_1, \dots, \gamma_6) = (2, 2, 2, 0, 0, 0)$. In the first case one uses the transformation constant $c = 5$, in the second case the transformation constant $c = 3$. When the θ -parameters are transformed accordingly one obtains two equivalent parameterizations. But in the first parameterization the second three items show DIF, in the second parameterization the first three items show DIF. It can not be decided which of the item sets shows DIF because both parameterizations are valid. The model builder fixes by the choice of the constraint which set of items shows DIF. But this basic identifiability problem seems worse than it is. When fitting a Rasch model one wants to identify the items that deviate from the model but assumes that the model basically holds for the majority of items. Thus one aims at identifying the maximal set of items for which the model holds. Thus, if, for example, the unconstrained items can be given by $(\gamma_1, \dots, \gamma_6) = (5, 5, 3, 3, 3, 2)$, the choice $\gamma_3 = 0$ makes

the items 3,4,5 Rasch-compatible and the rest has DIF. In contrast, $\gamma_6 = 0$ makes item 6 Rasch-compatible but the rest has DIF. Therefore, a natural choice is $\gamma_3 = 0$, where it should be emphasized again that any choice is legitimate. The fitting procedure proposed in the following will automatically identify the maximal set of items that is Rasch-compatible. We will come back to that in the following but give here general conditions for the identifiability of items.

In the general model with predictor $\eta_{pi} = \theta_p - \beta_i - \mathbf{x}_p^T \boldsymbol{\gamma}_i$ a set of identifiability conditions is

- (1) Set $\beta_I = 0$, $\boldsymbol{\gamma}_I^T = (0, \dots, 0)$ (or for any other item).
- (2) The matrix \mathbf{X} with rows $(1, \mathbf{x}_1^T), \dots, (1, \mathbf{x}_P^T)$ has full rank.

(for a proof, see Appendix B). The first condition means that for one item the β and the γ -parameters have to be fixed. It serves as a reference item in all populations. The second condition is a general condition that postulates that the explanatory variables have to contain enough information to obtain identifiable parameters. It is a similar condition as is needed in common regression models. It should be noted that the condition is general, the explanatory variables can be continuous or categorical. In the latter case, the matrix \mathbf{X} contains the dummy variables that code the categorical variable. As in regular regression, in particular highly correlated continuous covariates affect the rank of the design matrix and might yield unstable estimates. In the extreme case estimates are not unique because they are not identifiable. Then, one might reduce the set of covariates. In the case where estimates still exist but are unstable, nowadays regularization methods are in common use. A specific form of regularization is also used in the following.

4.3. Estimation by Regularization

4.3.1. Maximum Likelihood Estimation

Let the data be given by (Y_{pi}, \mathbf{x}_p) , $p = 1, \dots, P, i = 1, \dots, I$. Maximum likelihood estimation of the model is straightforward by embedding the model into the framework of generalized linear models. By using again the coding for persons and parameters in the parameter vectors $\mathbf{1}_{P(p)}$ and $\mathbf{1}_{I(i)}$ the model has the form

$$\begin{aligned} \log \left(\frac{P(Y_{pi} = 1)}{P(Y_{pi} = 0)} \right) &= \theta_p - \beta_i - \mathbf{x}_p^T \boldsymbol{\gamma}_i \\ &= \mathbf{1}_{P(p)}^T \boldsymbol{\theta} - \mathbf{1}_{I(i)}^T \boldsymbol{\beta} - \mathbf{x}_p^T \boldsymbol{\gamma}_i. \end{aligned}$$

With the total vector given by $(\boldsymbol{\theta}^T, \boldsymbol{\beta}^T, \boldsymbol{\gamma}_1^T, \dots, \boldsymbol{\gamma}_I^T)$ one obtains for observation Y_{pi} the design vector $(\mathbf{1}_{P(p)}^T, -\mathbf{1}_{I(i)}^T, 0, 0, \dots, -\mathbf{x}_p^T \dots, 0, 0)$, where the component $-\mathbf{x}_p^T$ corresponds to the parameter $\boldsymbol{\gamma}_i$.

Although ML estimation is straightforward estimates will exist only in very simple cases, for example, if the explanatory variable codes just two subgroups. In higher dimensional cases ML estimation will deteriorate and no estimates or selection of parameters are available.

4.3.2. Penalized Estimation

In the following we will consider regularization methods that are based on penalty terms. The general principle is, not to maximize the log-likelihood function, but a penalized version. Let $\boldsymbol{\alpha}$ denote the total vector of parameters, in our case $\boldsymbol{\alpha}^T = (\boldsymbol{\theta}^T, \boldsymbol{\beta}^T, \boldsymbol{\gamma}_1^T, \dots, \boldsymbol{\gamma}_I^T)$. Then one maximizes the penalized log-likelihood

$$l_p(\boldsymbol{\alpha}) = l(\boldsymbol{\alpha}) - \lambda J(\boldsymbol{\alpha}),$$

where $l(\cdot)$ is the common log-likelihood of the model and $J(\boldsymbol{\alpha})$ is a penalty term that penalizes specific structures in the parameter vector. The parameter λ is a tuning parameter that specifies how serious the penalty term has to be taken. A widely used penalty term in regression problems is $J(\boldsymbol{\alpha}) = \boldsymbol{\alpha}^T \boldsymbol{\alpha}$, that is, the squared length of the parameter vector. The resulting estimator is known under the name ridge estimate, see Hoerl and Kennard (1970) for linear models and Nyquist (1991), Segerstedt (1992), LeCessie (1992) for the use in GLMs. Of course, if $\lambda = 0$ maximization yields the ML estimate. If $\lambda > 0$ one obtains parameters that are shrunk toward zero. In the extreme case $\lambda \rightarrow \infty$ all parameters are set to zero. The ridge estimator with small $\lambda > 0$ stabilizes estimates but does not select parameters, which is the main objective here. Penalty terms that are useful because they enforce selection are L_1 -penalty terms.

Let us start with the simple case of a univariate explanatory variable, which, for example, codes gender. Then the proposed *lasso penalty for differential item functioning* (DIFlasso) is given by

$$J(\boldsymbol{\theta}^T, \boldsymbol{\beta}^T, \boldsymbol{\gamma}_1^T, \dots, \boldsymbol{\gamma}_I^T) = \sum_{i=1}^I |\gamma_i|,$$

which is a version of the L_1 -penalty or lasso (for least absolute shrinkage and selection operator). The lasso was propagated by Tibshirani (1996) for regression models, and has been studied intensively in the literature, see, for example, Fu (1998), Osborne et al. (2000), Knight and Fu (2000), Fan and Li (2001) and Park and Hastie (2007). It should be noted that the penalty term contains only the parameters that are responsible for differential item functioning, therefore only the parameters that carry the information on DIF are

penalized. Again, if $\lambda = 0$ maximization yields the full ML estimate. For very large λ all the γ -parameters are set to zero. Therefore, in the extreme case $\lambda \rightarrow \infty$ the Rasch model is fitted without allowing for differential item functioning. The interesting case is in between, when λ is finite and $\lambda > 0$. Then the penalty enforces selection. Typically, for fixed λ , some of the parameters are set to zero while others take values unequal zero. With a carefully chosen tuning parameter λ the parameters that yield estimates $\hat{\gamma}_i > 0$ are the ones that show DIF.

For illustration we consider a Rasch model with 10 items and 70 persons. Among the 10 items three suffer from DIF induced by a binary variable with parameters $\gamma_1 = 2$, $\gamma_2 = -1.5$, $\gamma_3 = -2$. Figure 4.1 shows the coefficient build-ups for the γ -parameters for one data set, that is, how the parameters evolve with decreasing tuning parameter λ . In this data

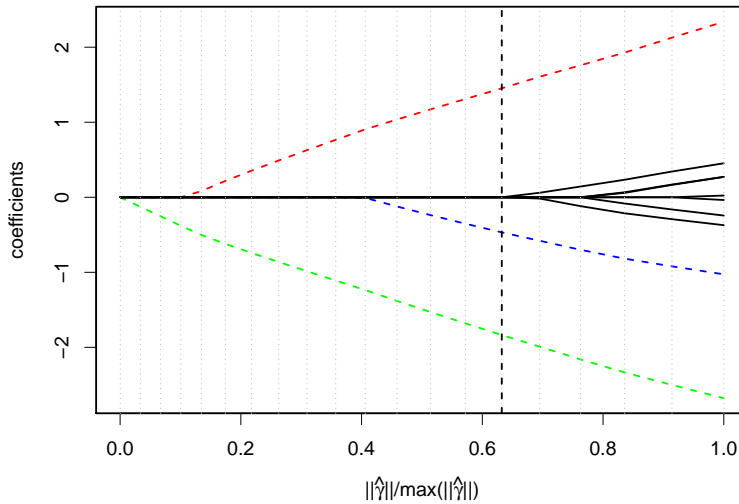


Figure 4.1.: Coefficient build-up for Rasch model with DIF induced by binary variable, dashed lines are the items with DIF, solid lines are the items without DIF.

set ML estimates existed. We do not use λ itself on the x -axis but a transformation of λ that has better scaling properties. Instead of giving the λ -values on the x -axis we scale it by $\|\hat{\gamma}\| / \max \|\hat{\gamma}\|$, where $\max \|\hat{\gamma}\|$ corresponds to the L_2 -norm of the maximal obtainable estimates, that is, the ML estimates. On the right side of Figure 4.1 one sees the estimates for $\lambda = 0$ ($\|\hat{\gamma}\| / \max \|\hat{\gamma}\| = 1$), which correspond to the ML estimates for the DIF model. At the left end all parameters are shrunk to zero, corresponding to the value of λ , where the simple Rasch model without DIF is fitted. Thus, the figure shows how estimates evolve over diminishing strength of regularization. At the right end no regularization is exerted, at the left side regularization is so strong that all γ -parameters are set to zero. The vertical line shows the tuning parameter selected by BIC (see below), which represents the best

estimate for this selection criterion. If one uses this criterion all items with DIF (dashed lines) are selected, obtaining estimates unequal zero. But for all items without DIF the estimates are zero. Therefore in this data set identification was perfect.

In the general case with a vector of covariates that potentially induce DIF a more appropriate penalty is a modification of the grouped lasso (Yuan and Lin, 2006; Meier et al., 2008). Let $\boldsymbol{\gamma}_i^T = (\gamma_{i1}, \dots, \gamma_{im})$ denote the vector of modifying parameters of item i , where m denotes the length of the person-specific covariates. Then the *group lasso penalty for item differential functioning* (DIFlasso) is

$$J(\boldsymbol{\theta}^T, \boldsymbol{\beta}^T, \boldsymbol{\gamma}_1^T, \dots, \boldsymbol{\gamma}_I^T) = \sum_{i=1}^I \|\boldsymbol{\gamma}_i\|,$$

where $\|\boldsymbol{\gamma}_i\| = (\gamma_{i1}^2 + \dots + \gamma_{im}^2)^{1/2}$ is the L_2 -norm of the parameters of the i th item with m denoting the length of the covariate vector. The penalty encourages sparsity in the sense that either $\hat{\boldsymbol{\gamma}}_i = \mathbf{0}$ or $\gamma_{ij} \neq 0$ for $j = 1, \dots, m$. Thus the whole group of parameters collected in $\boldsymbol{\gamma}_i$ is shrunk simultaneously toward zero. For a geometrical interpretation of the penalty, see Yuan and Lin (2006). The effect is that in a typical application only some of the parameters get estimates $\hat{\boldsymbol{\gamma}}_i \neq \mathbf{0}$. These correspond to items that show DIF.

Choice of Penalty Parameter

An important issue in penalized estimation is the choice of the tuning parameter λ . In our case it determines the numbers of items identified as inducing DIF. Therefore, it determines if all items with DIF are correctly identified and also if some are falsely diagnosed as DIF-items. To find the final estimate in the solution path it is necessary to balance the complexity of the model and the data fit. However, one problem is to determine the complexity of the model, which in penalized estimation approaches is not automatically identical to the number of parameters in the model. We worked with several criteria for the selection of the tuning parameter, including cross-validation and AIC criteria with the number of parameters determined by the degrees of freedom for the lasso (Zou et al., 2007). A criterion that yielded a satisfying balancing and which has been used in the simulations and applications is the BIC (Schwarz, 1978) with the degrees of freedom for the group lasso penalty determined by a method proposed by Yuan and Lin (2006). Here, the degrees of freedom (of penalized parameters $\boldsymbol{\gamma}$) are approximated by

$$\tilde{df}_{\boldsymbol{\gamma}}(\lambda) = \sum_{i=1}^I I(\|\boldsymbol{\gamma}_i(\lambda)\| > 0) + \sum_{i=1}^I \frac{\|\boldsymbol{\gamma}_i(\lambda)\|}{\|\boldsymbol{\gamma}_i^{ML}\|} (m - 1).$$

Since the person parameters and the item parameters are unpenalized, the total degrees of freedom are $df(\lambda) = I + P + \tilde{df}_\gamma(\lambda) - 1$. The corresponding BIC is determined by

$$BIC(\lambda) = -2 \cdot l(\boldsymbol{\alpha}) + df(\lambda) \cdot \log(P \cdot I),$$

where $l(\boldsymbol{\alpha})$ is the log-likelihood of the current parameter vector $\boldsymbol{\alpha}$.

Identifiability and Estimation

As demonstrated at the end of Section 4.2 the model without constraints is not identifiable. Moreover, the identification of DIF-items depends on the constraints that are used. Because of the basic identifiability problem, one can define few or many items as DIF-items. The aim to find the maximal set of Rasch-compatible items with a small set of items characterized as DIF-items is strongly supported by the regularization approach. We first fit the full model without constraints. Because of the regularization term the parameters are estimable, although not identifiable, see Friedman et al. (2010), where this procedure has been used in multinomial regression models. With growing smoothing parameter more and more items are characterized as not being compatible with the Rasch model with the items that have the strongest deviation from the Rasch model being the first ones that show in the coefficient build-ups. In all cases that were considered the value of the smoothing parameter chosen by our criterion was such that not all parameters showed DIF. Most often just few parameters had estimates $\hat{\gamma}_i \neq 0$. Therefore, one of the items with $\hat{\gamma}_i = 0$ is chosen and used as reference. By rearranging items, one of these items is denoted by I and one sets $\beta_I = 0$, $\gamma_I^T = (0, \dots, 0)$, which is obtained by computing $\hat{\beta}_i - \hat{\beta}_I$ for the item difficulties and $\hat{\gamma}_i - \hat{\gamma}_I$ for the γ -parameters, where $\hat{\beta}_i$, $\hat{\gamma}_i$ denote the estimates for the full model. This yields the identifiable parameters that are considered in the following simulations and applications. Of course, in the simulations the true values are centered around the same item.

Further Remarks

We focus on penalized ML estimation. Regularized estimation with penalty terms has the advantage that the penalty term is given explicitly and, therefore, it is known how estimates are shrunk. An alternative procedure that could be used is boosting as proposed in Chapter 5. It selects relevant variables by using weak learners and regularization is obtained by early stopping. Although the form of regularization is not given in an explicit form it typically is as efficient as regularization with corresponding penalty terms. Also mixed model methodology as used by Soares et al. (2009) to estimate DIF can be combined

with penalty terms that enforce selection. However, methodology is in its infancy, see for example Ni et al. (2010) or Bondell et al. (2010).

4.4. The Fitting Procedure At Work

In the present section it is investigated if the procedure is able to detect the DIF items. This is done in a simulation study where it is known which items are affected by DIF.

Illustration

For illustration, we will first consider several examples. In the first example we have 70 persons, 10 items, three with DIF ($\gamma_1^T = (-1, 0.8, 1)$, $\gamma_2^T = (-1.1, 0.5, 0.9)$, $\gamma_3^T = (1, -1, -1)$, $\gamma_4^T = \dots = \gamma_{10}^T = (0, 0, 0)$). The upper panel in Figure 4.2 shows the coefficient build-ups

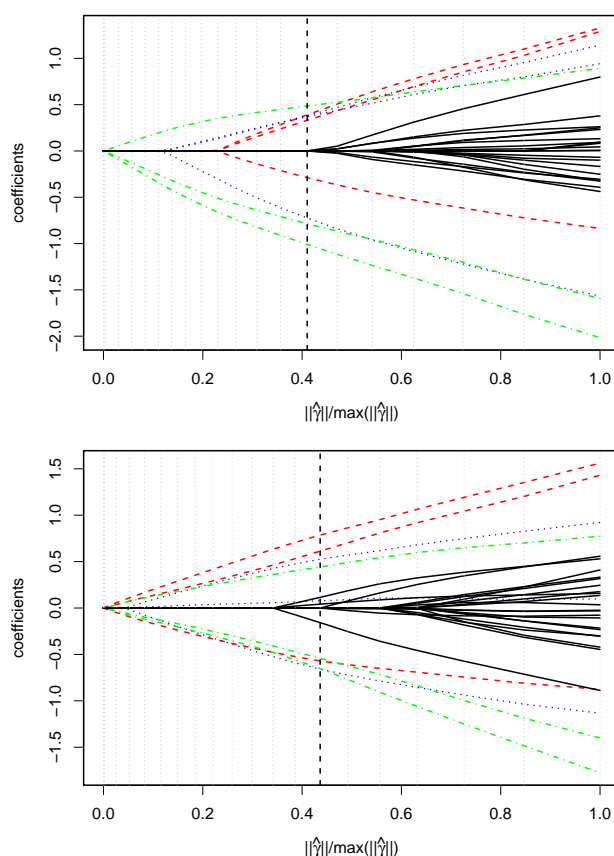


Figure 4.2.: Coefficient build-up for Rasch model with DIF induced by three variables, dashed lines are the items with DIF, solid lines are the items without DIF. Upper panel shows perfect identification, in the lower panel identification is not perfect.

for an exemplary data set. Now one item is represented by three lines, one for each co-variate. Again, items with DIF are given by non-solid lines and items with DIF by solid lines. In this data set the BIC criterion selects all the items with DIF and sets all items without DIF to zero. In the lower panel one sees a data set where identification is not perfect. It is seen that some items without DIF are falsely considered as inducing DIF. But also in this data set the items with DIF are the first ones to obtain estimates unequal zero when penalization is relaxed. The items without DIF obtain estimates unequal zero but estimates are very small.

An example without DIF is seen in Figure 4.3. The setting is the same as before ($P = 70$, $I = 10$) but all γ -parameters are set to zero. It is seen that the procedure also works well in the case of the Rasch model because all γ -parameters are estimated as zero.

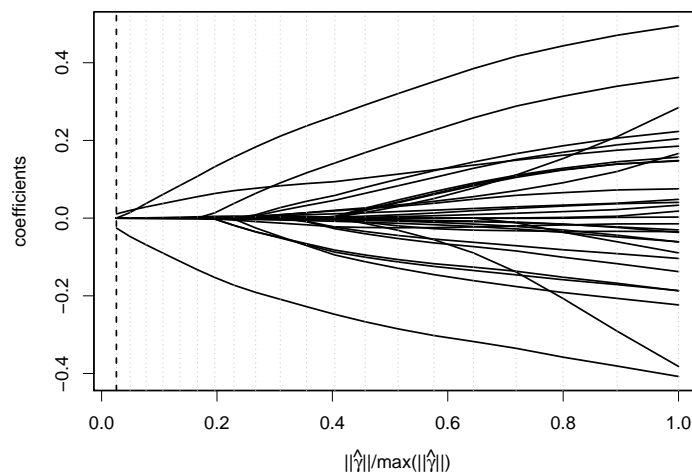


Figure 4.3.: Coefficient build-up for Rasch model without DIF .

For further illustration we show in the upper panel of Figure 4.4 the estimates of 100 simulated data sets for the same setting as in Figure 4.2. The boxplots show the variability of the estimates, the stars denote the underlying true values. The β -parameters in the left block represent the basic item parameter, which are estimated rather well. In the next block the modifying parameters γ_{is} are shown for items with DIF and in the last block the modifying parameters for items without DIF are shown. In this last block the stars that denote true values are omitted since they are all zero. Overall, the estimates of the basic β -parameters (first block) and the items without DIF (third block) are quite close to their true values. In particular the estimates of the parameters that correspond to items without DIF are zero or close to zero and are frequently diagnosed as not suffering from DIF. The γ -parameters in the middle block, which correspond to items with DIF, are distinctly unequal zero and, therefore, the DIF-items are identified. But the latter estimates are downward biased because of the exerted penalization, which shrinks the estimates.

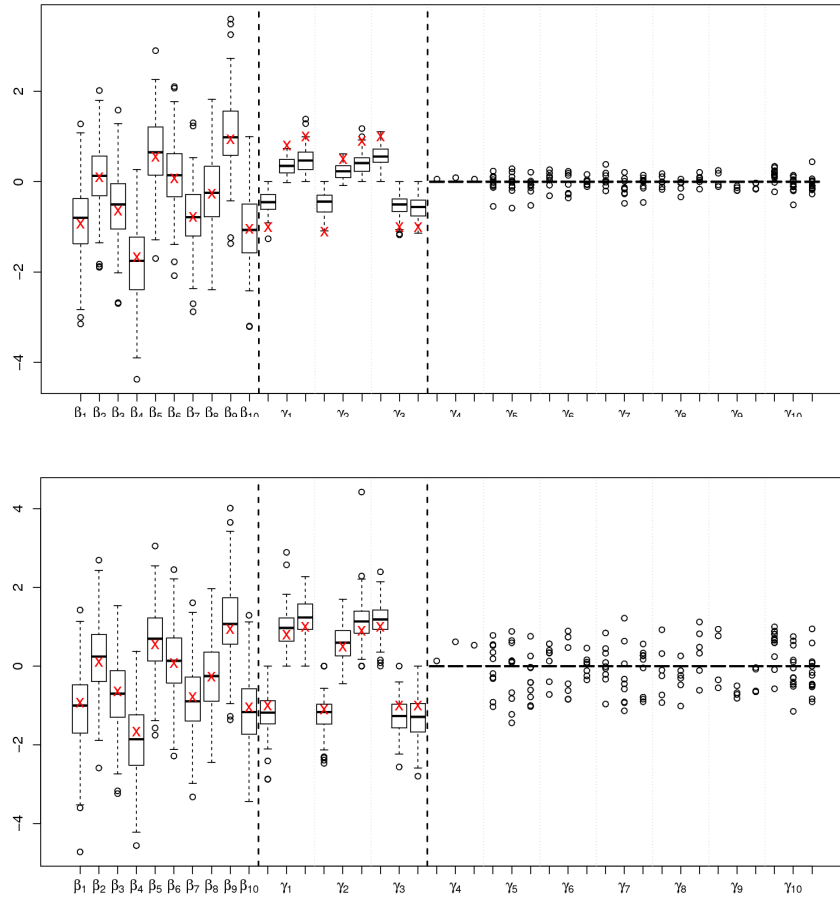


Figure 4.4.: Upper panel: Box plots of estimates for Rasch model with DIF induced by three variables, stars denote true values. Lower panel: the same model with a final ML step on selected items.

The bias can be removed and estimators possibly improved by an additional refit. The fit of the model in combination with the selection of the tuning parameter yields the set of items that are considered as suffering from DIF. To avoid shrinkage and bias one can compute a final un-penalized ML fit of the reduced model that contains only the parameters that have been selected as being non-zero. In the lower panel of Figure 4.4 the estimates with a final refit step are given. While the estimation of the basic β -parameters has hardly changed, the downward bias in item parameters for items with DIF is removed. However, the estimates of parameters for items without DIF automatically suffers. If one of these items is diagnosed as DIF-item the final ML-fit yields larger values than the penalized estimate. The reduction of bias comes with costs. As is seen from Figure 4.4 the variability for the procedure with an additional ML step is larger. Penalization methods like lasso typically have two effects, selection and shrinkage. By shrinking estimates extreme values are avoided and standard

errors are smaller but bias is introduced. The final ML estimate aims at a new balance of variance and bias but keeps the selection effect.

Simulation Scenarios

In the following we give results for selected simulation scenarios based on 100 simulations. The person parameters are drawn from a standard normal distribution and we consider scenarios with varying strength of DIF. The item parameters have the form $\beta_i + \mathbf{x}_p^T \boldsymbol{\gamma}_i$. We always work with standardized person characteristics \mathbf{x}_p , that is, the components have variance 1. A measure for the strength of DIF in an item is the variance $V_i = \text{var}(\beta_i + \mathbf{x}_p^T \boldsymbol{\gamma}_i)$, which for independent components has the simple form $V_i = \sum_j \gamma_{ij}^2$. For standardization it is divided by the number of covariates m . The average of $\frac{1}{m} \sqrt{V_i}$ over the items *with* DIF gives a measure of the strength of DIF in these items. The implicitly used reference value is the standard deviation of the person parameters, which is 1. We use three different strengths of DIF, strong, medium and weak. For the parameters of strong DIF, the DIF strength is 0.25. For medium and weak DIF, the parameters from the strong DIF setting are multiplied by 0.75 and 0.5, respectively. Accordingly, the DIF strengths for medium and weak are 0.1875 and 0.125. An overall measure of DIF in a setting is the average of $\frac{1}{m} \sqrt{V_i}$ over *all* items. For the strong scenario with 20 items one obtains 0.05, for the medium and weak 0.038 and 0.025, respectively.

When calculating mean squared errors we distinguish between person and item parameters. For person parameters it is the average over simulations of $\sum_p (\hat{\theta}_p - \theta_p)^2 / P$. For items it is the squared difference between the estimated item difficulty and the actual difficulty $\sum_p \sum_i [(\beta_i + \mathbf{x}_p^T \boldsymbol{\gamma}_i) - (\hat{\beta}_i + \mathbf{x}_p^T \hat{\boldsymbol{\gamma}}_i)]^2 / (I \cdot P)$.

One of the main objectives of the method is the identification of items with DIF. The criteria by which the performance of the procedure can be judged are the hits or true positives (i.e. the number of correctly identified items with DIF) and the false positives (i.e. the number of items without DIF that are falsely diagnosed as items with DIF).

The settings considered in the following are:

- Setting 1: 250 persons, 20 items, 4 with DIF on 5 variables, parameters (strong DIF): $\boldsymbol{\gamma}_1^T = (-0.8, 0.6, 0, 0, 0.8)$, $\boldsymbol{\gamma}_2^T = (0, 0.8, -0.7, 0, 0.7)$, $\boldsymbol{\gamma}_3^T = (0.6, 0, 0.8, -0.8, 0)$, $\boldsymbol{\gamma}_4^T = (0, 0, 0.8, 0.7, -0.5)$, $\boldsymbol{\gamma}_5^T = \dots = \boldsymbol{\gamma}_{20}^T = (0, 0, 0, 0, 0)$, two variables binary, three standard normally distributed.
- Setting 2: 500 persons, items as in setting 1,
- Setting 3: 500 persons, 20 items, 8 with DIF on 5 variables, items 1 – 4 as in setting 1, items 5 – 8 same as items 1 – 4

- Setting 4: 500 persons, 40 items, 8 with DIF, items 1 – 8 same as in setting 3
- Setting 5: same as Setting 2, but the person abilities differ along with the first (binary) covariate ($\theta|x_1 = 1 \sim N(1, 1)$, $\theta|x_1 = 0 \sim N(0, 1)$)

Settings 1-4 vary in the number of persons and the number of items with and without DIF. In all of them the person parameters are not linked to the predictor. After all, it can occur in practice that there is correlation between the abilities of persons and the grouping variable. Therefore it is of interest if the performance of DIF detection suffers from correlation. The last setting, setting 5, explicitly includes correlation between the abilities and the first, binary predictor. Persons with predictor value $x_1 = 1$ are assumed to have higher abilities.

In Table 4.1 the MSEs as well as the hits and false positive rates are given for the fit of the Rasch model (without allowing for DIF), the DIFlasso and the DIFlasso with refit. It is seen that the accuracy of the estimation of person parameters does not depend strongly on the strength of DIF. It is quite similar for strong and medium DIF and slightly worse for weak DIF. Also the fitting of the Rasch model or DIFlasso yields similar estimates of person parameters. The refit procedure, however, yields somewhat poorer estimates in terms of MSE. The estimation of item parameters shows a different picture. DIFlasso distinctly outperforms the Rasch model, in particular if DIF is strong the MSE is much smaller. The refit is better than the normal DIFlasso in all of the settings except one. Therefore, when the focus is on the estimation of item parameters, the refit can be recommended for a more precise and unbiased estimation.

The effect of correlation between abilities and predictors is investigated separately. The settings 2 and 5 use the same number of persons and parameters, but in setting 5 a binary covariate is highly correlated with the person abilities. The MSEs can be seen from Table 4.1. In Figure 4.5 the MSEs for the two settings are compared to each other for strong DIF with setting 2 being depicted in the left box plot and setting 5 in the right box plot. The upper panel shows the box plots for the MSEs of the person parameters, the lower for the item parameters. Again, it can be seen that for the person parameters the refit performs a little worse than the regular DIFlasso and the Rasch Model. The correlation in setting 5 makes estimation harder for all three methods but the estimation of person parameters does not suffer strongly. From the lower panel, which shows the MSEs for the item difficulties, it is seen that the DIFlasso strongly outperforms the Rasch Model and that the refit improves the estimation of the item-specific parameters. As for person parameters the correlation affects the accuracy of estimation but not very seriously. Estimation accuracy in terms of MSE does not suffer strongly from the presence of correlation.

Since our focus is on the identification of DIF-items the hits and false positive rates are of particular interest. It is seen from the lower panel of Table 4.1 that the procedure works

Setting		MSE person parameters			MSE item parameters		
		Rasch	DIFlasso	Refit	Rasch	DIFlasso	Refit
1	$P = 250$	strong	0.341	0.344	0.376	0.368	0.149
	$I = 20$	medium	0.349	0.350	0.370	0.233	0.145
	$I_{\text{DIF}} = 4$	weak	0.347	0.347	0.348	0.129	0.127
2	$P = 500$	strong	0.316	0.326	0.350	0.338	0.070
	$I = 20$	medium	0.323	0.328	0.345	0.202	0.064
	$I_{\text{DIF}} = 4$	weak	0.331	0.332	0.341	0.105	0.069
3	$P = 500$	strong	0.326	0.327	0.367	0.650	0.106
	$I = 20$	medium	0.327	0.328	0.358	0.378	0.096
	$I_{\text{DIF}} = 8$	weak	0.334	0.335	0.351	0.186	0.108
4	$P = 500$	strong	0.176	0.176	0.189	0.343	0.082
	$I = 40$	medium	0.176	0.176	0.185	0.205	0.071
	$I_{\text{DIF}} = 8$	weak	0.181	0.180	0.185	0.110	0.081
5*	$P = 500$	strong	0.333	0.342	0.366	0.355	0.091
	$I = 20$	medium	0.338	0.343	0.359	0.210	0.078
	$I_{\text{DIF}} = 4$	weak	0.345	0.346	0.353	0.109	0.082

Setting		true positives	false positives
1	$P = 250$	strong	0.99
	$I = 20$	medium	0.79
	$I_{\text{DIF}} = 4$	weak	0.04
2	$P = 500$	strong	1.00
	$I = 20$	medium	1.00
	$I_{\text{DIF}} = 4$	weak	0.71
3	$P = 500$	strong	1.00
	$I = 20$	medium	1.00
	$I_{\text{DIF}} = 8$	weak	0.77
4	$P = 500$	strong	1.00
	$I = 40$	medium	1.00
	$I_{\text{DIF}} = 8$	weak	0.61
5*	$P = 500$	strong	1.00
	$I = 20$	medium	0.99
	$I_{\text{DIF}} = 4$	weak	0.52

Table 4.1.: MSEs for the simulation scenarios (upper panel) and average rates of hits/false positives (lower panel).

* Setting 5 contains a binary covariate highly correlated with the person abilities.

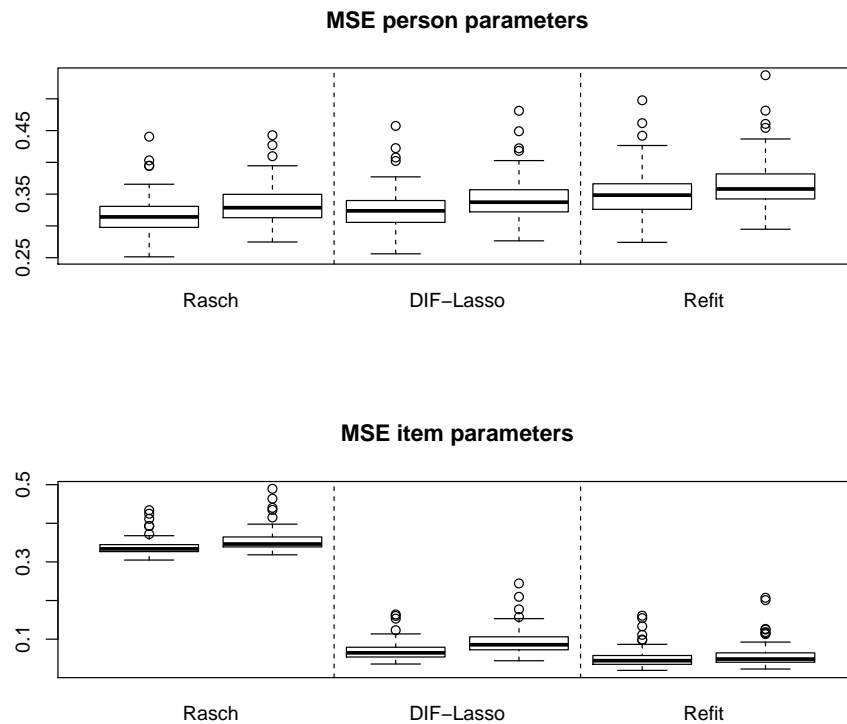


Figure 4.5.: Box plots of MSEs for setting 2 (left box plot) and setting 5 (right box plot) for strong DIF.

well. If DIF is strong the hit rate is 1 or close to 1, for medium DIF one needs more persons in the setting to obtain a hit rate of 1. Of course, for weak DIF identification is harder and one will not always find all the items with DIF. For 250 persons and weak DIF, there is not enough information anymore to have an acceptable selection performance. But the hit rate increases strongly when the number of persons is increased to 500 persons (setting 2) instead of 250 persons (setting 1).

One nice result is that the false positive rate is negligible. Although not all items with DIF may be found, it hardly occurs that items without DIF are falsely diagnosed. Only in setting 3 the false positive rate is slightly increased. When comparing the settings 2 and 5, which only differ because in the latter correlation between abilities and predictors is present, it is seen that the hit rate suffers only for weak DIF. For strong and medium DIF the performance is very similar. Together with the results for the MSEs, DIFlasso seems to perform rather well also in the case where the performance of persons is linked to a binary covariate. Differences in abilities and DIF are well separated.

Comparison with Methods for Multiple Groups

The method proposed here works for vector-valued predictors but can be compared to existing methods that are limited to the handling of groups. Most of the established methods for detection of uniform DIF use just two groups representing, for example, gender. Magis et al. (2010) set up a nice framework and shortly introduce into the existing DIF methods. For the case of one binary covariate they consider the Mantel-Haenszel (MH) method, developed by Mantel and Haenszel (1959) and applied to DIF by Holland and Thayer (1988), the method of logistic regression (Swaminathan and Rogers, 1990) and Lord's χ^2 -test (Lord, 1980). MH is a χ^2 -test where the performances of the groups are tested against each other separately for all items, conditional on the total test score. For the method of logistic regression, a logit model is fitted using the total test score, the group membership and an interaction of test score and group membership as covariates. The response is the probability of a person to score on an item. For detection of uniform DIF, the parameter for the group membership is tested by a likelihood ratio or a Wald test. Lord's χ^2 -test uses the null-hypothesis that the item parameters are equal within both groups. The parameters are estimated by the maximum likelihood principle separately for the groups, then they are tested against each other by a χ^2 -test. The methods can be generalized to the case of multiple groups. This has been done by Somes (1986) and Penfield (2001) for MH, Magis et al. (2011) for logistic regression and Kim et al. (1995) for Lord's χ^2 test. In R (R Core Team, 2015), these methods are implemented in the package `difR` (Magis et al., 2013), which is also described in Magis et al. (2010).

Since we are interested in the performance in the case of more complex predictors we give the results of a simulation study where DIF in more than two groups is investigated. For the comparison we use the implementation in `difR` (Magis et al., 2013). In the simulation study three different settings are considered. The definition of the DIF strengths strong, medium and weak is equivalent to the previous simulations. Each setting is run 100 times. We use $P = 500$ persons and $I = 20$ items. The groups are defined by a factor with q categories which is either $q = 5$ or $q = 6$. For the DIFlasso approach, this factor is represented by $q - 1$ binary dummy variables. For the reference methods, we have the case of a q -groups comparison. The number of DIF-items is either $I_{\text{DIF}} = 4$ or $I_{\text{DIF}} = 8$.

Table 4.2 shows the results for the selection performance of the single methods. It can be seen, that DIFlasso is competitive for strong and medium DIF. It achieves even lower false positive rates than the other methods. For weak DIF, however, the true positive rate is smaller than for the competing methods. It selects too few variables resulting in minimal false positive rates but too small true positive rates. The effect is ameliorated if the number of groups and the number of DIF items increases. It should also be noted that in the simple case of binary predictors the MH method and the other procedures designed explicitly for this case outperform the general method proposed here. Thus for few groups

Setting			DIFlasso	Lord	Logistic	MH
1	$P = 500$	strong	true positives	1.000	1.000	1.000
			false positives	0.016	0.020	0.058
	$I = 20$	medium	true positives	0.998	1.000	1.000
			false positives	0.009	0.018	0.052
	$I_{\text{DIF}} = 4$	weak	true positives	0.410	0.938	0.978
			false positives	0.000	0.018	0.056
2	$P = 500$	strong	true positives	1.000	1.000	1.000
			false positives	0.020	0.113	0.070
	$I = 20$	medium	true positives	1.000	1.000	1.000
			false positives	0.013	0.024	0.061
	$I_{\text{DIF}} = 4$	weak	true positives	0.890	0.985	0.995
			false positives	0.000	0.019	0.056
3	$P = 500$	strong	true positives	1.000	1.000	1.000
			false positives	0.053	0.042	0.118
	$I = 20$	medium	true positives	1.000	1.000	1.000
			false positives	0.028	0.028	0.093
	$I_{\text{DIF}} = 8$	weak	true positives	0.530	0.946	0.980
			false positives	0.001	0.021	0.064

Table 4.2.: Means of true positives and false positives for the simulations with multiple groups

and in particular for weak DIF the alternative methods are to be preferred. If the predictor structure is more complex the proposed method works well and allows to investigate the effect of vector-valued predictors.

Separating the Group Effects from the Abilities

In the following we briefly discuss how real differences in the populations in addition to DIF could be explicitly incorporated in a model. The main problem is that one has to model the effect of a grouping variable or, more general, a covariate on the ability of persons and still have an identifiable model. For categorical covariates, which are considered in the following, the separation of the group effect from DIF can be obtained by using an ANOVA-type representation of the model. But because of nesting the design is not that of a simple ANOVA model. Let the covariate be a categorical variable or factor like gender. Then one has two groups of persons, males and females. Because individuals have their own effects, individuals themselves can be seen as a factor. The third factor is determined by the items. A useful representation of the model treats gender, or, more general, the categorical covariate as a blocking factor. The individuals are the elements within a block, where it is essential that there is no connection between the individuals in different levels of

the blocking variable. For the representation, the index p for the individual is replaced by the index (g, j) , where g represents the level of the grouping variable ($g = 1, \dots, G$) and j the individuals within blocks ($j = 1, \dots, n_g$). There is no connection between observations (g, j) and (g', j) , $g \neq g'$, but between observations (g, j) and (g, j') because the latter are from the same level of the blocking variable. For a general treatment of nesting, see, for example, McCullagh and Nelder (1989).

The Rasch model (without DIF) for individual (g, j) and item i can then be represented by the predictor

$$\eta_{gji} = \eta_0 + \alpha_g + \delta_{gj} - \beta_i$$

with the usual symmetric side constraints $\sum_g \alpha_g = \sum_j \delta_{gj} = \sum_i \beta_i = 0$. In the model the person parameter θ_p has been replaced by the parameter $\eta_0 + \alpha_g + \delta_{gj}$, where (g, j) represents person p . The model contains the constant η_0 and three factors, the grouping variable, the persons, nested within groups, and the items. The parameter α_g represents the effect of the categorical covariate, which is separated from the effect δ_{gj} of person (g, j) . The model with DIF has the representation

$$\eta_{gji} = \eta_0 + \alpha_g + \delta_{gj} - (\beta_i + \gamma_{gi})$$

with the additional side constraints $\sum_g \gamma_{gi} = \sum_i \gamma_{gi} = 0$. The additional parameters γ_{gi} represent the interaction between the grouping variable and the items. It should be noted that the factor item interacts only with the grouping variable, not with the persons. This makes the model a very specific ANOVA-type model. Of course, alternative side constraints can be used. For example, $\sum_i \beta_i = 0$ can be replaced by $\beta_I = 0$, or $\sum_g \alpha_g = 0$ can be omitted if η_0 is fixed by $\eta_0 = 0$. Here we used symmetric side constraints because they are most often used in ANOVA-type models.

By using the embedding into the ANOVA framework with nesting structure one obtains an identifiable model that separates the effect of the grouping variable from the persons. It works also for more than one grouping variable by specifying main effects (and possibly interaction effects) of the grouping variables and nesting the persons within the blocks.

4.5. Examples

4.5.1. Exam Data

Our first data example deals with the solution of problems in an exam following a course on multivariate statistics. There were 18 problems to solve and 57 students. In this relatively small data set two variables that could induce DIF were available, the binary variables

level (bachelor student of statistics: 1, master student with a bachelor in an other area: 0) and gender (male: 0, female: 1). Figure 4.6 shows the coefficient build-ups. With BIC as selection criterion no item showed DIF. So we were happy that the results did not indicate that the exam was preferring specific subgroups.

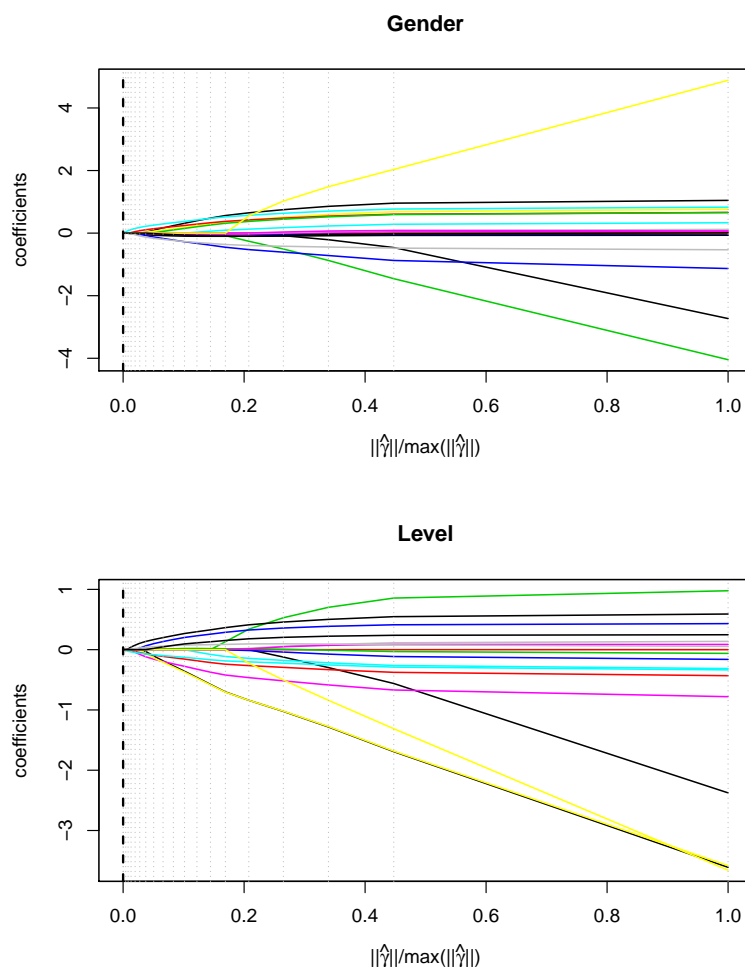


Figure 4.6.: Coefficient build-ups for exam data.

In this simple case, in which potential DIF is induced by binary variables, which indicate the sub populations, one can also use test statistics to examine if DIF is present because ML estimates exist. The embedding into the framework of generalized linear models allows to use the likelihood ratio test to test the null hypothesis $\gamma_1 = \dots, \gamma_I = 0$ (for the theory see, for example, Tutz (2012)). We consider the effects of gender and level separately. The p-values are 0.28 for gender and 0.38 for level. The result supports that DIF is not present. Alternatively, we used model checks based on conditional estimates as Andersen's likelihood ratio test (Andersen, 1973a), which is implemented in the R-package `eRm`, see Mair et al.

(2012) and Mair and Hatzinger (2007). These tests resulted in p-values of 0.315 for gender and 0.417 for level and also support that DIF is not an issue in this data set.

4.5.2. Knowledge Data

An example that has also been considered by Strobl et al. (2015) uses data from an online quiz for testing one's general knowledge conducted by the weekly German news magazine SPIEGEL. The 45 test questions were from five topics, politics, history, economy, culture, and natural sciences. We use the same sub sample as Strobl et al. (2015) consisting of 1075 university students from Bavaria, who had all been assigned a particular set of questions. The covariates that we included as potentially inducing DIF are gender, age, semester of university enrollment, an indicator for whether the student's university received elite status by the German excellence initiative (elite), and the frequency of accessing SPIEGEL's online magazine (spon).

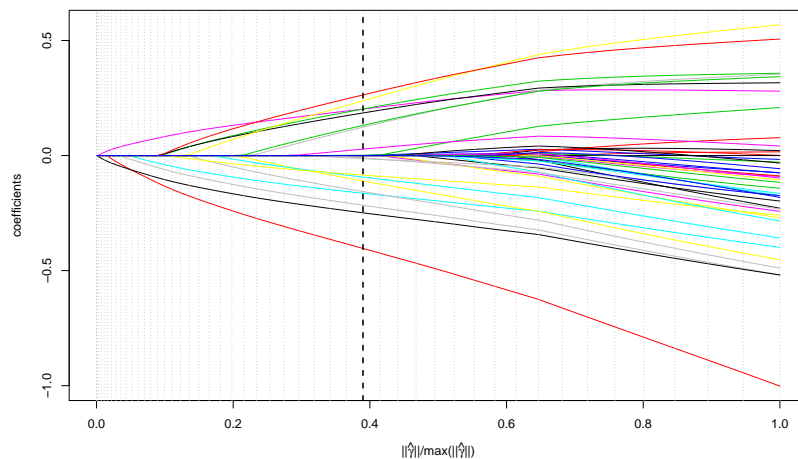


Figure 4.7.: Coefficient build-ups for covariate gender in Quiz Data; dashed vertical line indicates BIC-optimal path point

Figure 4.7 shows as an example the coefficient build-ups for the covariate gender. At the path point that was selected by the BIC criterion (dashed vertical line), 16 of the 45 items showed DIF, which is not surprising because it is not a carefully constructed test that really focusses on one latent dimension. In Figure 4.8, the estimated effects of the items containing DIF are visualized. The upper panel shows the profile plots of the parameters for the included covariates. For each item with DIF one profile is given. The lower panel shows the strengths of the effects in terms of the absolute value of the coefficients. One boxplot refers to the absolute values of the 16 parameters for one covariate. It is seen that the strongest effects are found for the covariate gender, the weakest effects are in the

variable elite, which measures the status of the university where the student is enrolled. It should be noted that the importance of the single covariates for the DIF can be measured by the absolute values of their coefficients since all covariates were standardized.

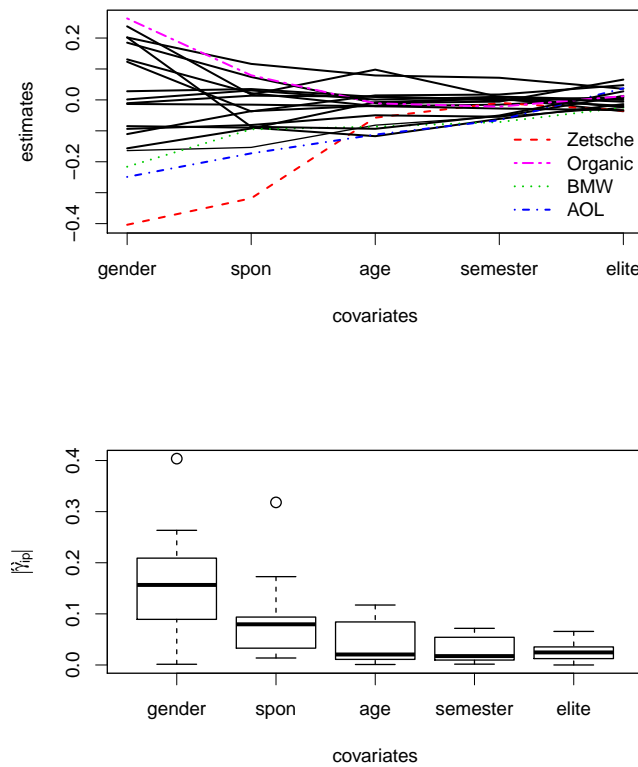


Figure 4.8.: Upper panel: profile plot for coefficient estimates of items with DIF, profiles of the four items with highest DIF are highlighted; lower panel: boxplots of absolute values of coefficient-estimates for items with DIF

In Figure 4.8 (upper panel) four items are represented by dashed lines. They showed the strongest DIF in terms of the L_2 -norm of the estimated parameter vector. All of them refer to economics. For illustration, these four items are considered in more detail. They are

- Zetsche: "Who is this?" (a picture of Dieter Zetsche, the CEO of the Daimler AG, maker of Mercedes cars, is shown).
- AOL: "Which internet company took over the media group Time Warner?"
- Organic: "What is the meaning of the hexagonal 'organic' logo?" (Synthetic pesticides are prohibited)
- BMW: "Which German company took over the British automobile manufacturers Rolls-Royce?"

The profiles for the items Zetsche, AOL and BMW are quite similar. They are distinctly easier for male participants and for frequent visitors of SPIEGELonline. The item Organic shows a quite different shape being definitely easier for females. It is also easier to solve for students that are not frequent visitors of SPIEGELonline. The item differs from the other three items because it refers more to a broad education than to current issues. Also, females might be more interested in (healthy) food in general. In this respect female students and students that do not follow the latest news seem to find the item easier. Therefore the different profile.

In Figure 4.9, the estimates for the covariate-specific parameters in DIF-items are illustrated using effect stars, see also Appendix A. In effect stars, one star represents the

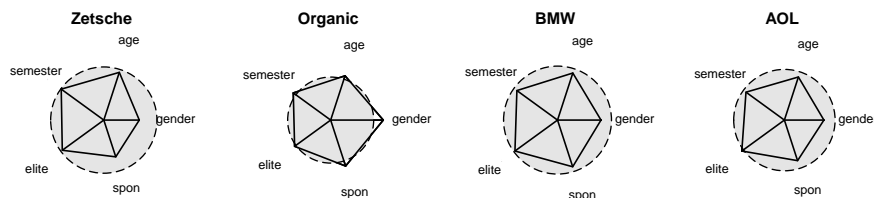


Figure 4.9.: Effect stars for the four items with highest DIF

parameters corresponding to one group of parameters. The length of the rays corresponds to the exponentials of the parameters. A circle with radius 1 represents the case of $\exp(0)$ and, therefore, the no-effect case. Rays within the circle represent negative parameters, rays beyond the circle represent positive parameters. In our application, all parameters corresponding to one item are collected in a star. It can be seen that among the presented items, the item organic is the only item with a positive gender effect.

4.6. DIFlasso with Variable Selection

The main objective of DIFlasso is to detect items containing DIF by regularization methods. After all, the proposed group-lasso type penalty term does not perform variable selection. For an item diagnosed as DIF item, every coefficient will be unequal zero, i.e. every covariate will have an effect. However, the results from the general knowledge test from Subsection 4.5.2 suggest that variable selection could be a desirable tool for our analysis. Clearly, the variables gender, age and spon seem to be important variables for the DIF items we found. The covariates semester and elite have rather small estimates for all items and

are likely not to induce any DIF. Therefore, an automatic method to select the variables actually inducing DIF could be useful.

For that purpose, two strategies seem sensible. The first possibility is to replace the group-lasso type penalty term by a simple lasso-type penalty. The respective penalty term can be denoted as

$$J(\boldsymbol{\theta}^T, \boldsymbol{\beta}^T, \boldsymbol{\gamma}_1^T, \dots, \boldsymbol{\gamma}_I^T) = \sum_{i=1}^I \sum_{j=1}^m |\gamma_{ij}|.$$

Every additional (compared to the basic Rasch model) item-specific parameter is penalized separately with respect to its absolute value. Therefore, every single of these parameters can be estimated as zero exactly and be eliminated from the model. If every parameter corresponding to a covariate j is eliminated from the model ($\gamma_{ij} = 0$ for $i = 1, \dots, I$), covariate j is completely eliminated from the model. In such a case, variable selection is realized. Yet, this concept is less focused on the issue of detecting DIF items. Every item i with at least one parameter unequal zero, i.e. $\hat{\gamma}_{ij} \neq 0$ holds for at least one j , $j = 1, \dots, m$, is diagnosed to be a DIF item.

The second possibility to perform variable selection is to extend the regular DIFlasso method by a post-selection step. The idea is to calculate another regularized model based on the items selected by DIFlasso. Instead of using the complete model (4.2), the reduced model

$$\log \left(\frac{P(Y_{pi} = 1)}{P(Y_{pi} = 0)} \right) = \begin{cases} \theta_p - (\beta_i + \mathbf{x}_p^T \boldsymbol{\gamma}_i) & i \in A \\ \theta_p - \beta_i & i \notin A \end{cases} \quad (4.3)$$

is used where A denotes the active set of items where DIF was found by DIFlasso. Therefore, item-specific parameters are used for DIF items only. To perform variable selection, we again use the regularization technique of group lasso. But, instead of grouping by items we group by covariates. The penalty term for the post-selection step in DIFlasso is therefore denoted as

$$J(\boldsymbol{\theta}^T, \boldsymbol{\beta}^T, \boldsymbol{\gamma}_1^T, \dots, \boldsymbol{\gamma}_{I_A}^T) = \sum_{j=1}^m \|\boldsymbol{\gamma}_{\cdot j}\|, \quad (4.4)$$

where $\boldsymbol{\gamma}_{\cdot j}$ denotes the vector of item specific parameters for the j -th covariate $\boldsymbol{\gamma}_{\cdot j} = (\gamma_{1j}, \dots, \gamma_{I_A j})$. I_A denotes the number of items in the active set A and m denotes the number of covariates. This penalty term allows for explicit variable selection. Model selection is again performed by model selection criteria, we again recommend the BIC. With decreasing penalty parameter, the covariate vectors enter the model as a whole. If the BIC-optimal penalty parameter is chosen large enough, some variables will be excluded from the model.

4.7. An Alternative Method

In contrast to most existing methods the proposed procedure allows to include all variables that might lead to DIF and identify the items with DIF. Quite recently Strobl et al. (2015) proposed a new procedure that is also able to investigate the effect of a set of variables. Therefore, it seems warranted to discuss the differences between our method and the recursive partitioning approach advocated by Strobl et al. (2015).

Recursive partitioning is similar to CARTs (Classification and Regression Trees), which were propagated by Breiman et al. (1984). For a more recent introduction see Hastie et al. (2009), or from a psychological viewpoint Strobl et al. (2009). The basic concept of recursive partitioning and tree methods in regression models is to recursively partition the covariate space such that the dependent variable is explained best. In the case of continuous predictors partitioning of the covariate space means that one considers splits in single predictors, that is, a predictor X is split into $X \leq c$ and $X > c$ where c is a fixed value. All values c are evaluated and the best split is retained. If a predictor is categorical splits refer to all possible subsets of categories. Recursive partitioning means that one finds the predictor together with the cut-off value c that explains the dependent variable best. Then given $X \leq c$ (and the corresponding sub sample) one repeats the procedure searching for the best predictor and cut-off value that works best for the sub sample with $X \leq c$. The same is done for the sub sample with $X > c$. The procedure of consecutive splitting can be visualized in a tree. Of course, there are many details to consider, for example, one has to define what best explanation of the dependent variable means, when to stop the procedure and other issues. For details see Breiman et al. (1984).

In item response models the partitioning refers to the predictors that characterize the persons. That means when using the person-specific variable X , for example, age, it is split into $X \leq c$ and $X > c$. The Rasch model is fit in these sub populations yielding different estimates of item parameters. Then one has to decide if the difference between item estimates before splitting and after splitting is systematic or random. If it is systematic the split is warranted. For the decision Strobl et al. (2015) use structural change tests, which have been used in econometrics (see also Zeileis et al. (2008)). Although the basic concept is the same as in the partitioning in regression models, now a model is fitted and therefore the method is referred to as model based partitioning. For details see Strobl et al. (2015).

For the knowledge data Strobl et al. (2015) identified gender, spon and age as variables that induce DIF. This is in accordance with our results (Figure 4.8), which also identified these variables as the relevant ones. By construction the partitioning approach yields areas, in which the effect is estimated as constant. The partitioning yielded eight subpopulations, for example, $\{female, spon \leq 1, age \leq 21\}$ and $\{male, spon \leq 2 - 3, age \leq 22\}$. Within these subspaces all items have estimates that are non-zero. Items that have particularly large

values are considered as showing DIF. It is not clear what criterion is used to identify the items that actually show DIF. Strobl et al. (2015) just describe 5 items that seem to have large values. Therefore, one can not compare the two approaches in terms of the number of selected items.

Let us make some remarks on the principles of the recursive partitioning approach to DIF and the penalization method proposed here.

Recursive partitioning can be considered a non-parametric approach as far as the predictors are concerned. No specific form of the influence of predictors on items is assumed. But, in the case of continuous variables implicitly a model is fitted that assumes that the effects are constant over a wide range, that is, over $X \leq c$ and $X > c$ given the previous splitting. In contrast, our penalization approach assumes a parametric model for DIF. Although it can be extended to a model with unspecified functional form, in the present version it is parametric. An advantage of parametric models is that the essential information is contained in a modest number of parameters that show which variables are influential for specific items. A disadvantage of any parametric model is that it can be misspecified. The partitioning approach, considered as a more exploratory tool, is less restrictive, although assuming a constant value over wide ranges is also a restriction.

An advantage of the parametric model, if it is a fair approximation to the underlying structure, is the use of familiar forms of the predictor, namely a linear predictor, which, of course, can include interactions. In contrast, partitioning methods strongly focus on interactions. Typically in each consecutive layer of the tree a different variable is used in splitting. The result is smaller and smaller subpopulations which are characterized as a combination of predictors. The subpopulations $\{female, spon \leq 1, age \leq 21\}$ and $\{male, spon \leq 2 - 3, age \leq 22\}$, found for the knowledge data seem rather specific.

A potential disadvantage of tree based methods is their instability. A small change of data might result in quite different splits. That is the reason why tree-based methods have been extended to random trees, which are a combination of several trees on the same data set, see Breiman (2001).

The penalty approach uses an explicit model for DIF, and the model is separated from the estimation procedure. In the partitioning approach the model and the fitting are entwined. For practitioners it is often helpful to have an explicit form of the model that shows how parameters determine the modelled structure. Moreover, in the penalty approach an explicit criterion is used to determine how many and which items show DIF. The ability to identify the right items has been evaluated in the previous section.

Of course, none of the models is true. Neither is the effect constant within an interval of age as assumed in the partitioning approach nor is the effect linear as assumed in the suggested model. But, as attributed to Box, although all models are wrong some can be useful. Since

the models are not nested a goodness-of-fit tests could yield a decision. But goodness-of-fit as a measure for the adequacy of a model is a tricky business in partitioning models as well as in regularized estimation procedures, in particular in the framework of item response models. Therefore, not much is available in terms of goodness-of-fit, although it might be an interesting topic of future research.

One basic difference seems to be that the penalty approach uses all covariates, with the variables that are of minor relevance obtaining small estimates, but selects items. The partitioning approach selects variables, or, more concisely combinations of covariates, but then estimates all items as having an effect, that is, estimates are unequal zero. Thus penalty approaches focus on the selection of items, partitioning methods on the selection of combinations of covariates.

4.8. Concluding Remarks

A general model for DIF that is induced by a set of variables is proposed and estimation procedures are given. It is shown that the method is well able to identify items with DIF. The concept is general, with modifications it can be extended to models that include items with more than two categories as, for example, the graded response model (Samejima, 1997) or the partial credit model (Masters, 1982). Also the assumption that items are modified in the linear form $\mathbf{x}_p^\top \boldsymbol{\gamma}_i$ can be relaxed to allow for additive functions $f_1(x_{p1}) + \dots + f_m(x_{pm})$ by using, for example, P-spline methodology (Eilers and Marx, 1996).

The estimation used here is penalized unconditional ML estimation. Alternative regularized estimators could be investigated, for example, estimators based on mixed models methodology. Also the regularization technique can be modified by using boosting techniques instead of penalization.

The method is implemented in the R package `DIFlasso` (Schauberger, 2014a) and is available from CRAN. It uses the the coordinate ascent algorithm proposed in Meier et al. (2008) and the corresponding R package `grplasso` (Meier, 2009).

5. Detection of Differential Item Functioning in Rasch Models by Boosting Techniques

5.1. Introduction

In the beginnings of item response theory (IRT) the focus was on the Rasch model (Rasch, 1960) and its extensions to the 2PL and 3PL model by Birnbaum (1968). The Rasch model assumes that every person has a fixed latent ability and every item has a fixed difficulty. The difference between ability and difficulty determines the probability that a person solves an item. The extensions from Birnbaum (1968) attenuated this assumption by introducing two additional item parameters, for discrimination and guessing. Since then, item response theory has been a topic of intensive research and has been extended in various ways.

A well-known problem in item response models is that the probability to score on an item might vary over persons with the same latent ability. This may be caused by certain characteristics of the persons like gender, age, or race or by other unknown (latent) classes within the tested population. The phenomenon is known under the name Differential Item Functioning (DIF). If an item is detected to have DIF one option is to remove the item because it does not provide a fair measurement of the respective trait.

There is a wide range of literature on DIF in general, see, for example, Holland and Wainer (2012) and Millsap and Everson (1993). A very popular choice to detect DIF is the Mantel-Haenszel (MH) method. It is based on a test statistic proposed by Mantel and Haenszel (1959) and was used to detect DIF in item response theory by Holland and Thayer (1988). Various other methods to identify items which induce DIF have been proposed, for example, Swaminathan and Rogers (1990) and Lord (1980). Magis et al. (2010) set up a framework for the existing DIF methods and gave an excellent overview on currently available methods along with a software implementation (Magis et al., 2013).

This chapter is a modified version of Schauburger and Tutz (2015b), previous work on the issue can be found in the conference paper Schauburger and Tutz (2014). See Chapter 1 for more information on the personal contributions of all authors and textual matches.

The essential drawback of the MH method and most of the other existing methods is that they are limited to identify DIF between two subgroups, e.g., for male and female participants. Some methods for multiple subgroups have been developed, see *Somes (1986)*, *Penfield (2001)*, *Magis et al. (2011)* and *Kim et al. (1995)*. *Gonçalves et al. (2013)* set up a quite general Bayesian multifactor model for the detection of DIF in the 3PL model (*Birnbaum, 1968*). But methods that are able to handle DIF induced by continuous covariates or by a whole vector of covariates at the same time are scarce. Recently, *Strobl et al. (2015)* proposed to use tree methodology whereas *Magis et al. (2015)* used penalization techniques. The proposed method is strongly related to the approach proposed in Chapter 4 using regularization techniques.

In this Chapter, a new and efficient method is proposed for the detection of DIF in Rasch models that can deal with several (continuous and categorical) covariates and also interactions between the covariates simultaneously. The method is based on boosting techniques which have been developed more recently in the machine learning community (*Freund et al., 1996*) and in statistics (*Bühlmann and Hothorn, 2007a*), but their potential has not yet been exploited to uncover structures in item response models.

In Section 5.2, a DIF model is given in which DIF is explicitly represented by parameters. Section 5.3 introduces the idea of boosting in general, whereas Section 5.4 describes in detail the proposed estimation algorithm. Sections 5.5 and 5.6 illustrate the method by applications to both simulated and real data sets and compare it to existing approaches.

5.2. Differential Item Functioning Model

In the binary Rasch model the probability for a person to score on an item is determined by a parameter for the latent ability of the person and a parameter for the item difficulty. In the case of P persons and I items, the Rasch model is given by

$$P(Y_{pi} = 1) = \frac{\exp(\theta_p - \beta_i)}{1 + \exp(\theta_p - \beta_i)} \quad p = 1, \dots, P, \quad i = 1, \dots, I, \quad (5.1)$$

where Y_{pi} represents the response of person p on item i . It is coded by $Y_{pi} = 1$ if person p solves item i and $Y_{pi} = 0$ otherwise. Both the person parameters, θ_p , $p = 1, \dots, P$, and the item parameters, β_i , $i = 1, \dots, I$, are unknown and have to be estimated. Alternatively, model (5.1) can be given in the form

$$\log \left(\frac{P(Y_{pi} = 1)}{P(Y_{pi} = 0)} \right) = \theta_p - \beta_i, \quad (5.2)$$

where the left hand side specifies the so-called log-odds or logits. As model (5.2) is not identifiable in this general form, a restriction on the parameters is needed. A common choice, that is also used in the following, is $\theta_P = 0$. Alternatively, also one item parameter or the sum of all item parameters could be restricted to zero.

In item response models, DIF appears if an item has different difficulties depending on characteristics of the person which tries to solve the item. Therefore, DIF changes the item difficulty depending on covariates of the participants. This concept can be formalized by

$$\log \left(\frac{P(Y_{pi} = 1)}{P(Y_{pi} = 0)} \right) = \theta_p - (\beta_i + \mathbf{x}_p^T \boldsymbol{\gamma}_i), \quad (5.3)$$

where $\mathbf{x}_p^T = (x_{p1}, \dots, x_{pm})$ denotes a person-specific covariate vector of length m and, again, the restriction $\theta_P = 0$ is used. This general Differential Item Functioning Model (DIF model) is an extension of the Rasch model (5.2) allowing for person-specific item difficulties $\beta_i + \mathbf{x}_p^T \boldsymbol{\gamma}_i$. The item-specific parameters $\boldsymbol{\gamma}_i^T = (\gamma_{i1}, \dots, \gamma_{im})$ determine how the covariates x_{p1}, \dots, x_{pm} influence the difficulty of item i for person p . The original Rasch model corresponds to the special case where $\boldsymbol{\gamma}_i = \mathbf{0}$ for all items. The general model (5.3) was proposed in Chapter 4, a special case of the model was considered by Paek and Wilson (2011). However, estimation methods were quite different from the approach suggested here.

The main problem with the general DIF model is that $m \cdot I$ additional parameters (compared to the Rasch model) have to be estimated. Since each item has its own parameter per covariate, the number of parameters in the model can be huge. As the full DIF model is not identifiable, Maximum likelihood (ML) estimation is no option in this case. One possibility to overcome this problem are penalization methods where a penalized likelihood is maximized. For example, the ridge estimator (Hoerl and Kennard, 1970) or the lasso estimator (Tibshirani, 1996) can still be calculated when regular ML estimation fails. In Chapter 4 penalization methods of this type were used. Here we propose a quite different method, namely boosting. Boosting is an algorithmic procedure with origins in machine learning, see, for example, Freund and Schapire (1997).

Boosting as a method of statistical learning was developed by Friedman et al. (2000) and extended, for example, by Bühlmann and Yu (2003), Tutz and Binder (2006), Bühlmann (2006) and Bühlmann and Hothorn (2007a). Boosting in basic regression methods is available for the statistical software R (R Core Team, 2015), which will be used for all following calculations. It is, for example, implemented in the add-on package `mboost`, see Hothorn et al. (2013), which is also used for our computations.

One strength of boosting is, that it is able to select relevant terms in the predictor even in very high dimensional settings. This establishes the link to DIF in item response models.

The general assumption for our model is that only some of the items show DIF and only for these items item-specific parameters γ_i have to be estimated. Therefore, detection of DIF means selection of variables, or, in parametric models, selection of parameters that should be included in the model and, therefore, have estimates unequal zero. If a whole vector γ_i is set to zero, the difficulty of item i does not depend on the covariates and no DIF is present.

Generally, in the following all covariates are assumed to be standardized. This has the advantage that the covariates have the same scale and, therefore, can be compared directly. Especially, estimates for the item-specific covariates γ_{ip} can be compared directly and represent the size of the respective DIF effect.

5.3. Basic Boosting Procedures

Before developing boosting procedures for DIF models, in this section we briefly consider the basic concept of boosting and the choice of tuning parameters. The adaptation to DIF models will be considered in the consecutive sections. We start with the linear model, where boosting is much easier to conceptualize, and then proceed to boosting for generalized linear model (GLM).

Let us first consider a linear regression model

$$y_i = \beta_0 + \sum_{j=1}^p x_{ij}\beta_j + \epsilon_i, \quad i = 1, \dots, n,$$

for p covariates. If p is very large and it is suspected that not all covariates are influential, maximum likelihood estimation is a bad choice because of its instability in high-dimensional settings. In contrast, boosting is able to fit additive structures even in high-dimensional settings by successively fitting only parts of the model.

A basic ingredient of boosting is the specification of the so-called base learners $\hat{f}(\cdot)$. The base learners specify the structure that is fitted within one step of the procedure. Since we want to fit a linear model, the base learners are the ordinary least squares (OLS) estimators

$$\hat{f}(x) = \hat{\beta}_s x_s$$

for *single* covariates $s = 1, \dots, p$. That means within one step only one covariate is used. Although all the updates of all covariates are evaluated, in each boosting step a specific covariate s^* is selected that yields the greatest reduction of the residual sum of squares given the previous estimate. Let $\hat{\eta}^{(l-1)}$ represent the current linear predictor (the current model fit) from the previous boosting step $l - 1$, then the residual of the i th observation

is $u_i = y_i - \hat{\eta}^{(l-1)}$. In the l th step one fits a linear model for only one covariate on the data (u_i, x_{is}) , $i = 1, \dots, n$, and then selects the best predictor, which is used to update the linear predictor. Thus boosting is a stepwise procedure, which iteratively improves the fit by fitting a model to the current residuals. Tukey (1977) proposed the so-called "twicing", which means, after fitting a model one fits it again on the residuals. Boosting means not only two but many iterative fits. The following algorithm can be seen as the basic boosting procedure for linear models.

L_2 Boost in Linear Models

Step 1 (Initialization)

Given data $\{y_i, \mathbf{x}_i\}$, fit the base procedure to yield the function estimate $\eta^{(0)}(\mathbf{x}_i)$. Typically one fits an intercept model obtaining $\eta^{(0)}(\mathbf{x}_i) = \hat{\beta}_0$.

Step 2 (Iteration: Fitting of base learners and selection)

For $l = 1, 2, 3, \dots$, compute the residuals $u_i = y_i - \hat{\eta}^{(l-1)}(\mathbf{x}_i)$ and fit the base learners to the current data $\{u_i, \mathbf{x}_i\}$.

(a) One fits by minimizing least squares, that is, for fixed j one minimizes

$$\sum_{i=1}^n (u_i - \beta_j x_{ij})^2,$$

obtaining $\hat{\beta}_j$. (b) Selection means that one determines s^* such that

$$s^* = \arg \min_j \sum_{i=1}^n (u_i - \hat{\beta}_j x_{ij})^2.$$

(c) The improved fit is obtained by the update

$$\hat{\eta}^{(l)}(\mathbf{x}_i) = \hat{\eta}^{(l-1)}(\mathbf{x}_i) + \nu \hat{\beta}_{s^*} x_{is^*}.$$

Step 3 (Stop)

Iterate *Step 2* until $l = l_{stop}$ is reached.

In step 2, the linear predictor is updated by $\hat{\eta}^{(l)}(\mathbf{x}_i) = \hat{\eta}^{(l-1)}(\mathbf{x}_i) + \nu \hat{\beta}_{s^*} x_{is^*}$, which serves to compute the residuals in the following boosting step. It should be noted that the linear structure is maintained and only one component of the linear predictor is updated. Let $\hat{\eta}^{(l-1)}(\mathbf{x}_i)$ have the linear form $\sum_{j=1}^p \hat{\beta}_j^{(l-1)} x_{ij}$, then the addition of $\nu \hat{\beta}_{s^*} x_{is^*}$ changes only the weight on the variable s^* . The parameter ν , $0 < \nu \leq 1$, is used as a shrinkage parameter.

This shrinkage parameter makes the base learners "weak" and, therefore, prevents overfitting because only small steps towards the optimal solution are made. For this purpose, ν has to be chosen sufficiently small, $\nu = 0.1$ is a common choice. The procedure corresponds to a stepwise fitting of the linear model. In every step, one of the coefficients is updated by a rather small amount. The weakness of the learner is important because only then the fit is efficient (Bühlmann and Yu (2003) and Bühlmann (2006)). The smaller ν is chosen, the weaker the learner gets but the more boosting steps are required. If one does not stop, boosting is a complicated way of obtaining the maximum likelihood estimate. The selection effect is obtained by stopping the procedure before it converges. Then, only the variables that obtained non-zero weights are included in the model and one obtains a regularized estimate. Bühlmann (2006) showed that the procedure is consistent for underlying regression functions that are sparse in terms of the L_1 -norm.

Boosting can also be seen as a stepwise optimization of a specific loss function. For the linear regression model, the optimized loss function is the L_2 loss between the response and the linear predictor. In this context, boosting can be seen as a gradient descent method and sometimes is called *gradient boosting*. For the (slightly modified) L_2 loss function

$$L(y, \eta) = \frac{1}{2}(y - \eta)^2,$$

the gradient is given by the residuals

$$\frac{\partial L(y, \eta)}{\partial \eta} = y - \eta.$$

Therefore, instead of stepwise fitting of the residuals boosting can be seen as repeated fitting of the response with a so called offset, which is a known constant. In our case it is given by the estimate of the previous step $\hat{\eta}^{(l-1)}(\mathbf{x}_i)$. The least squares estimate uses the criterion $\sum_{i=1}^n (u_i - \beta_j x_{ij})^2 = \sum_{i=1}^n (y_i - (\hat{\eta}^{(l-1)}(\mathbf{x}_i) + \beta_j x_{ij}))^2$. In the latter form it is seen that one minimizes the least squares criterion for the original data y_i , but including the known constant $\hat{\eta}^{(l-1)}(\mathbf{x}_i)$ in the fit.

The iterative fitting with an offset offers a way to obtain boosting estimates also for generalized linear models (GLM). A GLM is in particular determined by the structure $\mu_i = E(y_i | \mathbf{x}_i) = h(\eta_i)$, where $h()$ is a known response function and the linear predictor has the form $\eta_i = \sum_{j=1}^p \beta_j x_{ij}$. One difference between the L_2 boost and a generalized linear model boosting is that in the boosting step one cannot fit a GLM to the residuals because, for example, with binary data, residuals are not from $\{0, 1\}$. The role of the residuals is taken by the offset.

Typically, the boosting algorithm is repeated for a large predefined number of steps l_{stop} . After the end of the algorithm, an appropriate criterion is used to determine the optimal

number of steps l_{opt} . This can either be done by information criteria like AIC or BIC or by the method of cross validation. For the example of the linear model, this corresponds to a model selection between l_{stop} possible models. The first model simply represents a null model where no covariates are included. With every boosting step, a new covariate is added or (if the respective covariate has been selected before) the parameter of a covariate is updated. As the base learners are assumed to be “weak”, successive models only differ slightly from each other. This makes it more likely for the optimal model to be found. Implicitly, this model selection corresponds to variable selection. Typically, in the finally chosen model l_{opt} , not all of the possible predictors have been chosen and, therefore, are excluded from the final model. Thus, l_{opt} is the most important regularization parameter for the boosting algorithm. A quite different approach to bypass the problem of overfitting is stability selection, which is described in detail in a following section and which will be applied to our DIFboost algorithm.

5.4. Boosting in Differential Item Functioning

5.4.1. The DIF Model as a Generalized Linear Model

The Rasch model and also the more general DIF model (5.3) can be embedded into the framework of generalized linear models (GLM).

Let the data be given by (Y_{pi}, \mathbf{x}_p) , $p = 1, \dots, P$, $i = 1, \dots, I$. For simplicity, we use the notation $\mathbf{1}_{P(p)}^T = (0, \dots, 0, 1, 0, \dots, 0)$ and $\mathbf{1}_{I(i)}^T = (0, \dots, 0, 1, 0, \dots, 0)$, where $\mathbf{1}_{P(p)}$ and $\mathbf{1}_{I(i)}$ have lengths $P - 1$ and I and have the value 1 at positions p and i , respectively. Therefore, the vectors are constructed in a way that they can be seen as dummy variables for the corresponding persons and items, respectively. Then, model (5.3) can be represented as

$$\begin{aligned} \log \left(\frac{P(Y_{pi} = 1)}{P(Y_{pi} = 0)} \right) &= \theta_p - \beta_i - \mathbf{x}_p^T \boldsymbol{\gamma}_i \\ &= \mathbf{1}_{P(p)}^T \boldsymbol{\theta} - \mathbf{1}_{I(i)}^T \boldsymbol{\beta} - \mathbf{x}_p^T \boldsymbol{\gamma}_i = \mathbf{z}_{pi}^T \boldsymbol{\alpha}. \end{aligned} \quad (5.4)$$

Here, $\boldsymbol{\alpha}^T = (\boldsymbol{\theta}^T, \boldsymbol{\beta}^T, \boldsymbol{\gamma}_1^T, \dots, \boldsymbol{\gamma}_I^T)$ denotes the complete parameter vector containing $\boldsymbol{\theta}^T = (\theta_1, \dots, \theta_{P-1})$ and $\boldsymbol{\beta}^T = (\beta_1, \dots, \beta_I)$. The design vector for the person p and the item i is denoted by $\mathbf{z}_{pi}^T = (\mathbf{1}_{P(p)}^T, -\mathbf{1}_{I(i)}^T, 0, \dots, 0, -\mathbf{x}_p^T, 0, \dots, 0)$. In \mathbf{z}_{pi} , the position of the component $-\mathbf{x}_p$ corresponds to the parameter $\boldsymbol{\gamma}_i$ in $\boldsymbol{\alpha}$.

In general, model (5.4) represents the structural component of a GLM for binary response with logit link. GLMs are extensively investigated in McCullagh and Nelder (1989), introductions with the focus on categorical data are found in Agresti (2002) and Tutz (2012).

Of course, also the regular Rasch model can be represented in the GLM framework by

$$\log \left(\frac{P(Y_{pi} = 1)}{P(Y_{pi} = 0)} \right) = \theta_p - \beta_i = \mathbf{1}_{P(p)}^T \boldsymbol{\theta} - \mathbf{1}_{I(i)}^T \boldsymbol{\beta}, \quad (5.5)$$

where the design vector and the parameter vector reduce to $(\mathbf{1}_{P(p)}, -\mathbf{1}_{I(i)})$ and $(\boldsymbol{\theta}^T, \boldsymbol{\beta}^T)$, respectively.

5.4.2. The DIFboost Algorithm

The objective of our approach is to detect DIF by boosting the logistic DIF model (5.3). Because selection refers to DIF-effects only, it is sensible to start the boosting selection procedure after the basic Rasch model has been fitted. The initial step is to fit the regular Rasch model (5.5). The result from this step are parameter estimates for the person parameters and the item parameters. The model fit from this first step is used as starting point for the further steps where boosting techniques are used to select potential DIF-effects. A similar approach was used by Boulesteix and Hothorn (2010) in a quite different context. In the following, our algorithm is described in detail.

The starting point for the algorithm is to fit a regular Rasch model to our data. This is done by embedding the Rasch model into the logistic regression model (5.5). It can be estimated by standard software as, for example, the function `glm` for the statistical software R (R Core Team, 2015). Then, one obtains estimates $\hat{\boldsymbol{\theta}}^T = (\hat{\theta}_1, \dots, \hat{\theta}_{P-1})$ for the person parameters and $\hat{\boldsymbol{\beta}}^T = (\hat{\beta}_1, \dots, \hat{\beta}_I)$ for the item difficulties. For a single observation, a linear predictor $\hat{\eta}_{pi} = \hat{\theta}_p - \hat{\beta}_i$ can be calculated which can be used to predict the probability of person p to score on item i to be $P(Y_{pi} = 1) = \frac{\exp(\hat{\eta}_{pi})}{1 + \exp(\hat{\eta}_{pi})}$. The linear predictors from the Rasch model for all person-item combinations are collected in $\hat{\boldsymbol{\eta}}_{\text{RM}} = (\hat{\eta}_{11}, \hat{\eta}_{12}, \dots, \hat{\eta}_{IP})$ and are passed on to the further steps of the algorithm.

For the boosting steps, the Rasch model (5.2) is extended to the more general DIF model (5.3). The parameters of the DIF model determine the base learners that are used. In our case, the model consists of three components, namely the person parameters, the item parameters and the item-specific covariate parameters. Therefore, each of these components serves as a possible base learner:

$$\tilde{\eta}(\mathbf{x}_p, p, i) = \begin{cases} \tilde{\theta}_p, & p = 1, \dots, P-1 \\ \tilde{\beta}_i, & i = 1, \dots, I \\ \mathbf{x}_p^T \tilde{\boldsymbol{\gamma}}_i, & i = 1, \dots, I \end{cases} \quad (5.6)$$

It is noteworthy that all base learners are linear. Nevertheless, they refer to different types of components that contain differing numbers of parameters (e.g. $\tilde{\gamma}_i$ vs. $\tilde{\beta}_i$). In cases like this, it is essential to ensure that all base learners share the same complexity so that the chances to be chosen are balanced. The complexity of base learners is determined by their degrees of freedom, which can be adapted by using internal penalty terms. In the case of linear base learners typically ridge penalties are used. Therefore, all the base learners presented above are restricted to have one degree of freedom by applying a ridge penalty when fitting the model. For more details on the complexity of base learners, see Hofner et al. (2011).

In every boosting step, only one of the base learner is updated, namely the one which yields the strongest reduction of an adequate loss function. The loss function that is used,

$$L(Y_{pi}, \tilde{\pi}_{pi}) = -(Y_{pi} \log(\tilde{\pi}_{pi}) + (1 - Y_{pi}) \log(1 - \tilde{\pi}_{pi})), \quad (5.7)$$

is the negative log-likelihood of a logit model with binary response. For boosting step l , this can be denoted by

$$\tilde{\eta}^*(\mathbf{x}_p, p, i) = \underset{\tilde{\theta}_p, \tilde{\beta}_i, \mathbf{x}^T \tilde{\gamma}_i}{\operatorname{argmin}} \sum_{p,i} L(Y_{pi}, \tilde{\pi}_{pi})$$

where the fitted probability $\tilde{\pi}_{pi}$ is calculated by fitting the model

$$\tilde{\pi}_{pi} = \frac{\exp(\tilde{\eta}^{(l)})}{\exp(1 + \exp(\tilde{\eta}^{(l)}))} \text{ with predictor } \tilde{\eta}^{(l)} = \tilde{\eta}^{(l-1)} + \tilde{\eta}(\mathbf{x}_p, p, i),$$

separately for every base learner from (5.6).

The estimates for the single candidates of the base learner are obtained by fitting logit models where the linear predictor from the current model fit is used as known offset and the respective base learner is the only predictor. Therefore, based on the current model fit, in every step only the base learner with the highest gain of information is updated. An additional parameter ν , $0 < \nu < 1$, regulates the step size of the parameter updates. It is chosen sufficiently small (typically $\nu = 0.1$) and only allows for small changes in every step. The parameter ν makes the base learners “weak” and is used to prevent quick overfitting. This procedure is repeated for a predefined number of steps l_{stop} .

For the first boosting step, the offset is chosen to be the linear predictor $\hat{\eta}_{RM}$ from the Rasch model, $\tilde{\eta}^{(0)} = \hat{\eta}_{RM}$. This provides two advantages: First, the person parameters θ and item parameters β are, in contrast to the item-specific covariate parameters γ_i , essential for the interpretability of model (5.3). Therefore, it is sensible to prevent those parameters from being excluded from the model. From this point of view, the offset provides starting values for the person and item parameters. Second, the object of our approach is to detect the improvement of the model fit by extending the Rasch model to the DIF model. Therefore,

we start from the model fit of the regular Rasch model. The boosting steps (possibly) add the information from the covariates. At some point during the boosting procedure, it can become necessary to adapt the person or the item parameters. Consequently, they can also be chosen as base learners within the boosting algorithm.

Typically, the model fitted after l_{stop} steps is overfitted and, therefore, not desirable. Two different strategies exist to finally identify the optimal model. One possibility is early stopping. Here, an optimal boosting step l_{opt} has to be found, either by an information criterion or by cross-validation. By early stopping, the boosting algorithm has the desirable effect of variable selection. The final model will only contain some of the possible parameters from model (5.3), namely the ones that have at least once been found to be the best base learner before the optimal step l_{opt} . The second option is stability selection, which will be discussed in detail later. For our analysis, we tried both early stopping using the BIC criterion and stability selection with similar results. As stability selection provided slightly more stable results, the option of early stopping is omitted for the rest of the chapter.

DIFboost algorithm

In the following, the outlined DIFboost algorithm is shortly sketched:

DIFboost

Step 1 (Initialization)

- Fit (5.5) for given scores Y_{pi} and initialize the offset $\tilde{\eta}^{(0)} = \hat{\eta}_{RM}$.
- Initialize $\tilde{\theta}_p = 0$, $p = 1, \dots, P - 1$, $\tilde{\beta}_i = 0$ and $\tilde{\gamma}_i = \mathbf{0}$, $i = 1, \dots, I$
- Set $l = 0$

Step 2 (Iteration)

- $l \rightarrow l + 1$
- Fit a logit model for every possible base learner where $\tilde{\eta}^{(l-1)}$ is used as offset
- Select the best base learner $\eta^*(\mathbf{x}_p, p, i)$
- Update the linear predictor by

$$\tilde{\eta}^{(l)} = \tilde{\eta}^{(l-1)} + \nu \tilde{\eta}^*(\mathbf{x}_p, p, i)$$

Step 3 (Stop)

Iterate *Step 2* until $l = l_{stop}$ is reached.

5.4.3. Illustrating Example

For illustration, first a single simulated data set will be considered. The data set is randomly drawn from Setting 2 (medium) of the simulation study in Section 5.5.2. We have $P = 500$ persons, $I = 20$ items (4 items with DIF, 16 without DIF) and $m = 5$ covariates, $l_{stop} = 500$ boosting steps are performed.

Figure 5.1 shows the coefficient paths along the boosting steps from $l = 0$ to $l = l_{stop} = 500$.

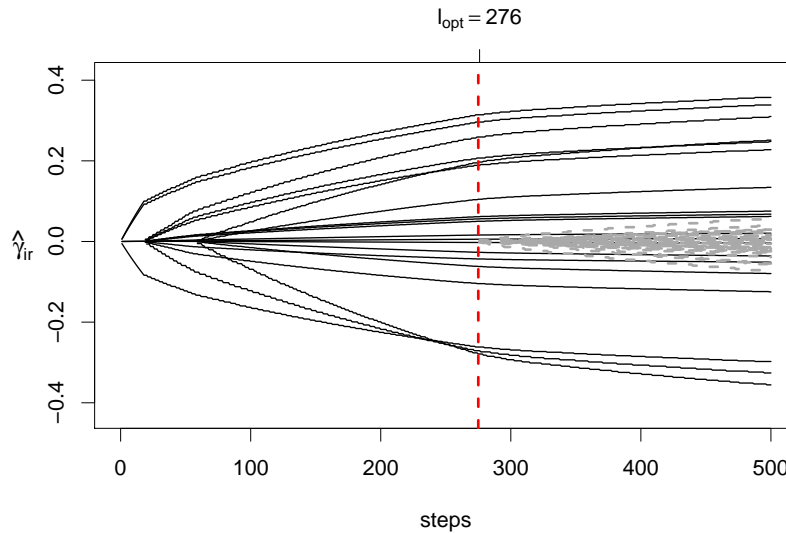


Figure 5.1.: Boosting paths of item-specific parameters $\hat{\gamma}_{ir}$ for exemplary data set; solid paths represent DIF items, dashed paths represent non-DIF items; dashed vertical line represents theoretically optimal boosting step l_{opt}

The solid black lines represent the paths of the four DIF items, the DIF-free items are represented by dashed gray lines. Every item is represented by five paths because $m = 5$ covariates are used to find DIF. This makes the plot hard to digest as it is hard to distinguish between the different items. Figure 5.2 reduces the plot to one path per item. Here, a path represents the Euclidean norm of the item-specific parameter vectors γ_i of the corresponding item i . This plot is much clearer and easier to interpret than Figure 5.1 although some information is suppressed. The DIF items (black solid lines) can clearly be separated from the other items (dashed gray lines) because they are updated much earlier in the boosting algorithm and, therefore, seem to be much more informative for the response. The dashed

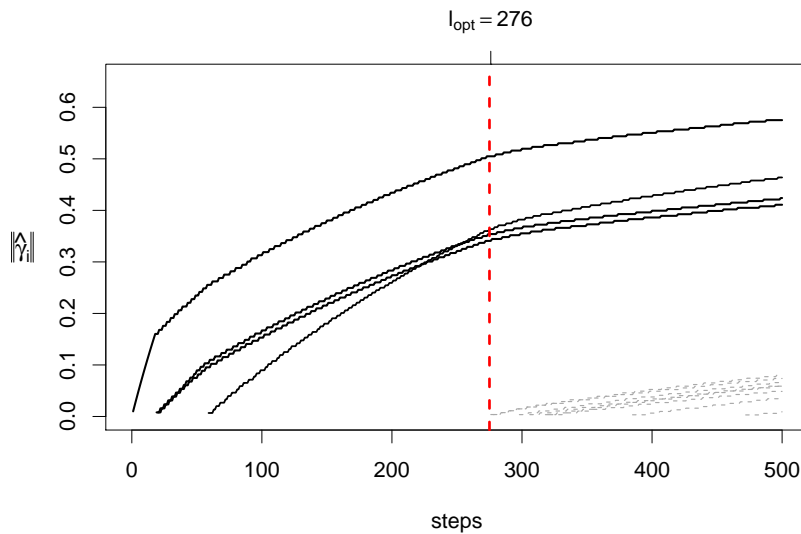


Figure 5.2.: Boosting paths of Euclidian norms of item-specific parameter vectors $\hat{\gamma}_i$ for exemplary data set; solid paths represent DIF items, dashed paths represent non-DIF items; dashed vertical line represents theoretically optimal boosting step l_{opt}

vertical line represents the theoretically optimal model where all DIF items are in the model and all DIF-free items are excluded.

5.4.4. Stability Selection

Choosing the optimal number of boosting steps via the BIC (or any other information criterion) has some drawbacks and may, therefore, not always be the best choice. One drawback is that the variable selection implied by the BIC can be unstable. Variables (or more precisely base learners) only have to be chosen in one single boosting step to be part of the final model. Therefore, it may happen that some items are diagnosed to have DIF although they have minimal coefficient estimates. This could lead to an increased false positive rate. Another drawback is that information criteria such as the BIC or the AIC use the degrees of freedom. Bühlmann and Hothorn (2007a) or Hofner et al. (2011). For example, the degrees of freedom can be determined using the hat matrix of the boosting algorithm, as proposed by Bühlmann and Hothorn (2007a) and Hofner et al. (2011). Yet, this is very time-consuming and also led to controversial methodological discussions, see Hastie (2007) and Bühlmann and Hothorn (2007b).

These drawbacks can be avoided by the concept of stability selection which was developed by Meinshausen and Bühlmann (2010). It is a very general approach which can be applied to a broad range of methods that include variable selection. It is based on the common

idea of model/variable selection by subsampling. This can be computationally beneficial because it allows for parallelized computations. Furthermore, it addresses the problem of unstable variable selection by pooling over many subsamples.

For the DIF model (5.3), stability selection can be obtained in the following way: For a predefined number of replications B , $\lfloor \frac{P}{2} \rfloor$ persons are drawn randomly from the original data set. The data set for one replication consists only of the observations in this subsample of persons. For each of the subsamples, the boosting algorithm is executed until l_{stop} . Then, one counts how often a specific base learner was selected at each specific step $l = 0, \dots, l_{stop}$. This gives the probabilities $\hat{\Pi}_i^l$ (or rather the relative frequencies over the B replications) of the base learner i to be in the model at a specific boosting step l . The probabilities are illustrated by so-called stability paths along the boosting steps as displayed in Figure 5.3. Finally, all base learners are selected with stability paths beyond a certain threshold value. These base learners represent the most frequent elements within the selected active set and, therefore, have to be considered as influential. In our application, we want to know which items have DIF and, therefore, we are only interested in the stability paths for γ_i for all items.

Stability selection is mainly determined by two parameters. The first parameter is q , which denotes how many distinct base learners are taken into the model when boosting the subsamples. As soon as q base learners have been selected, the procedure is stopped for the respective subsample. If less than q base learners are selected at $l = l_{stop}$, l_{stop} has to be increased. In the following, we choose 60% as a reasonable upper bound of the percentage of DIF-Items within a test and, therefore, $q = 0.6 \cdot I$. The second parameter is the threshold value for the selection probabilities of the single base learners which is denoted by π_0 . It is used to finally determine the set \hat{S}^{stable} of stable base learners. This set is defined by

$$\hat{S}^{stable} = \left\{ i : \max_{l=1, \dots, l_{stop}} (\hat{\Pi}_i^l) \geq \pi_0 \right\}.$$

According to Meinshausen and Bühlmann (2010), the threshold value should be chosen within a range of $\pi_0 \in (0.6, 0.9)$, also depending on the choice of q and the desired sparseness of the final model.

Although two parameters have to be determined in advance, stability selection proved to be very stable. Especially the choice of q turned out not to be crucial as long as it is chosen in a reasonable range. The main tuning parameter of the procedure is the threshold parameter π_0 . In our analysis, $\pi_0 = 0.9$ turned out to be a good choice. The threshold parameter π_0 is comparable to the level of significance in test-based procedures. In the simulation studies presented in the following section, $\pi_0 = 0.9$ caused false positive rates of about 5% if no DIF was present which is a popular choice for the level of significance in test-based procedures.

We use stability selection as a method of variable selection, but it does not provide parameter estimates. Estimates for the identified DIF effects are obtained by fitting a final DIF model for the selected items by maximum likelihood estimation. For illustration, Figure 5.3 shows the stability paths for the simulated data set from subsection 5.4.3 where four out of 20 items have DIF. We used $q = 0.6 \cdot I = 12$ and $B = 500$ subsamples. The stability paths for the 4 DIF items are drawn with solid lines. They can clearly be separated from the stability paths of the DIF-free items which are drawn with dashed lines. The threshold value $\pi_0 = 0.9$ is depicted by a dashed horizontal line. With the given threshold value, all DIF items are identified, all DIF-free items are not selected.

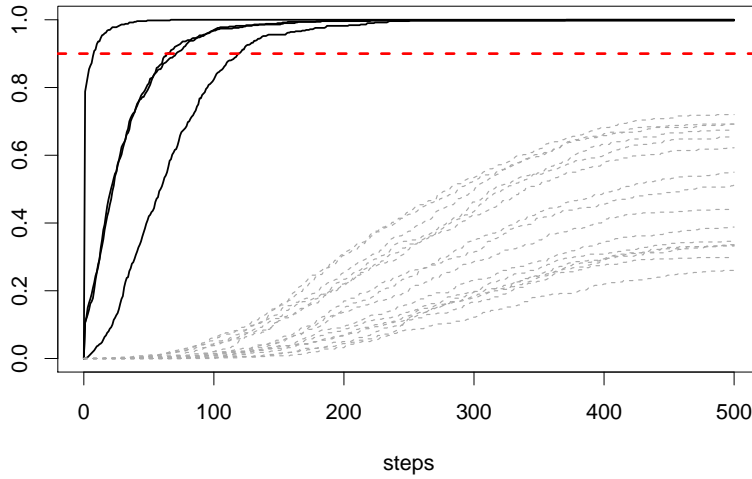


Figure 5.3.: Stability paths for exemplary data set; solid paths represent DIF items, dashed paths represent non-DIF items; dashed horizontal line represents threshold values $\pi_0 = 0.9$

5.4.5. Identifiability

Without any further constraints, the DIF-Model (5.3) is not identifiable. If person p tries to solve item i , let the linear predictor be denoted by $\eta_{pi} = \theta_p - \beta_i - \mathbf{x}_p^T \boldsymbol{\gamma}_i$. We set $\theta_P = 0$, which is a common constraint to obtain identifiability in simple Rasch models. However, in the DIF model a fixed vector \mathbf{c} allows to reparameterize the linear predictor to obtain

$$\eta_{pi} = \theta_p - \beta_i - \mathbf{x}_p^T \boldsymbol{\gamma}_i = \underbrace{\theta_p - \mathbf{x}_p^T \mathbf{c}}_{\tilde{\theta}_p} - \beta_i - \mathbf{x}_p^T \underbrace{(\boldsymbol{\gamma}_i - \mathbf{c})}_{\tilde{\boldsymbol{\gamma}}_i}.$$

Thus, the parameter sets $\{\theta_p, \beta_i, \boldsymbol{\gamma}_i\}$ and $\{\tilde{\theta}_p, \beta_i, \tilde{\boldsymbol{\gamma}}_i\}$ describe the identical model. This identification problem could be solved by restricting at least one item (the so-called reference

item R) to have parameters $\gamma_R = \mathbf{0}$. But, by definition this item can not have DIF and, therefore, would have to be chosen carefully. In particular, the choice of the reference item (or the corresponding \mathbf{c}) determines how many items show DIF (see also Chapter 4). A sensible strategy is to select the constraints in a way that only few items show DIF. In this respect the boosting approach offers a natural solution. The starting point of the algorithm is the Rasch model and, therefore, the best model fit if no DIF is permitted. Step by step, the DIF parameters are updated. During the boosting algorithm, every item which has not yet been chosen as a DIF item can be used as reference item. Therefore, the models are identifiable as long as at least one item is left out. In practice, one of the left-out items is chosen to be the reference item R and for reasons of simplicity, we then use the additional restriction $\beta_R = 0$ instead of $\theta_P = 0$.

5.5. Simulation Study

A simulation study is performed to illustrate the performance of the method in terms of identification of DIF items. First, the method is compared to established methods of DIF detection. This is done by simulation settings with only one binary or multi-categorical covariate which can also be handled by existing methods. The second part of the simulation will deal with settings with several (both continuous and categorical) covariates. These settings can not be compared directly to established methods and are compared to the recently published approach of DIFlasso from Chapter 4.

5.5.1. Comparison to Established Methods

Methods

Typically, in publications DIF is considered only for two groups, namely reference group and focal group. The standard method for this purpose is the Mantel-Haenszel (MH) method proposed by Holland and Thayer (1988). The method consists in computing a χ^2 -test that compares the performances of the groups separately for all items, conditional on the total test score.

Alternative methods are, among others, Lord's χ^2 -test (Lord, 1980) and the logistic regression method (Swaminathan and Rogers, 1990). In Lord's χ^2 -test, for each group the

parameters are estimated separately. Afterwards, a χ^2 -test is used that tests the null hypothesis of equal item parameters for both groups. The logistic regression method for the detection of uniform DIF uses the model

$$\log \left(\frac{P(Y_{pi} = 1)}{P(Y_{pi} = 0)} \right) = \beta_0 + \beta_1 s_p + \beta_2 x_p \quad (5.8)$$

for every item i , where s_p is the total test score of person p and x_p encodes the group membership. A test on uniform DIF is performed by a likelihood ratio test ($\alpha = 0.05$) on the null hypothesis $H_0 : \beta_2 = 0$. Model (5.8) can be extended by including a parameter for the interaction between the total test score and the group membership. This parameter could be used to test for non-uniform DIF. After all, as the focus of this chapter is on uniform DIF, this extension will not be considered here.

For the more general case of multi-group comparisons, the presented methods have been extended by Somes (1986) and Penfield (2001) for MH, Kim et al. (1995) for Lord's χ^2 test, and Magis et al. (2011) for logistic regression.

All results from the present chapter, including this simulation study, have been conducted by the statistical software R (R Core Team, 2015). The three reference methods for the simulation study are implemented in the add-on package `difR`, see Magis et al. (2010) and Magis et al. (2013). The level of significance was chosen to be $\alpha = 0.05$ for all performed tests.

Settings

The simulation study encompasses five different settings. Each setting is performed for different strengths of DIF, where the strength is measured by

$$\frac{1}{I_{\text{DIF}}} \sum_{i=1}^{I_{\text{DIF}}} \left(\frac{1}{m} \sqrt{\sum_{j=1}^m \gamma_{ij}^2} \right),$$

and I_{DIF} encodes the number of DIF-items. The term $\sum_{j=1}^m \gamma_{ij}^2$ represents the variance of the item difficulties $\beta_i + \mathbf{x}_p^T \boldsymbol{\gamma}_i$ for standardized covariates, where m again encodes the number of covariates. Therefore, the DIF strength in the simulations is measured as the mean of the variance of the item difficulties while accounting for the number of covariates. For details on measuring the DIF strength, see Chapter 4. The DIF strength in the simulation varies between 0.3 (very strong), 0.15 (strong), 0.1125 (medium) and 0.075 (weak).

For each setting, $P = 500$ persons and $I = 20$ items were generated, abilities θ and difficulties β were drawn from standard normal distributions. The number of groups and

the number of DIF items are varied. In the following, we present the five used settings and the used parameters for ‘strong’ DIF, for a different DIF strength the parameters are simply multiplied by a suited factor.

Setting 1 $I_{\text{DIF}} = 4$ DIF items, $k = 2$ groups, $\gamma_1 = 0.15$, $\gamma_2 = -0.15$, $\gamma_3 = 0.1$, $\gamma_4 = -0.2$, $\gamma_5, \dots, \gamma_{20} = 0$

Setting 2 $I_{\text{DIF}} = 8$ DIF items, $k = 2$ groups, $\gamma_1 = \gamma_5 = 0.15$, $\gamma_2 = \gamma_6 = -0.15$, $\gamma_3 = \gamma_7 = 0.1$, $\gamma_4 = \gamma_8 = -0.2$, $\gamma_9, \dots, \gamma_{20} = 0$

Setting 3 same as **Setting 1**, but the abilities are highly correlated with the group membership: $\theta_i|x_i = 0 \sim N(0, 1)$, $\theta_i|x_i = 1 \sim N(1, 1)$

Setting 4 $I_{\text{DIF}} = 4$ DIF items, $k = 5$ groups, $\gamma_1 = (0.4, 0, 0.3, -0.3)$, $\gamma_2 = (0.5, 0.4, -0.2, 0)$, $\gamma_3 = (0, -0.2, 0.4, 0.3)$, $\gamma_4 = (-0.2, 0.4, 0, 0.4)$, $\gamma_5 = \dots = \gamma_{20} = (0, 0, 0, 0)$

Setting 5 $I_{\text{DIF}} = 8$ DIF items, $k = 5$ groups, $\gamma_1 = \gamma_5 = (0.4, 0, 0.3, -0.3)$, $\gamma_2 = \gamma_6 = (0.5, 0.4, -0.2, 0)$, $\gamma_3 = \gamma_7 = (0, -0.2, 0.4, 0.3)$, $\gamma_4 = \gamma_8 = (-0.2, 0.4, 0, 0.4)$, $\gamma_9 = \dots = \gamma_{20} = (0, 0, 0, 0)$

In addition, the general settings 1,3, and 4 were run under the assumption that no DIF is present ($I_{\text{DIF}} = 0$). The only difference between the corresponding settings 1 and 3 is that in setting 3 the abilities correlate with the group membership.

Results

For every setting, 100 replications were performed. Table 5.1 shows the results for DIFboost ($q = 12$ and $\pi_0 = 0.9$) and the three reference methods in terms of true positive rate (TPR) and false positive rate (FPR). The true positive rate is determined by the rate of correctly identified DIF items. Therefore, higher values represent better performance. The false positive rate represents the rate of DIF-free items which have been assigned to be DIF items by mistake. Higher values represent worse performance.

For weak or medium DIF in settings 1-3, DIFboost outperforms MH and Lord in terms of TPR with similar FPR. Logistic regression shows both higher TPR and FPR. For strong and very strong DIF, DIFboost shows lower FPR than the competitors. In the multi-group settings 4-5, DIFboost again shows very low FPR but also partly lower TPR. All in all, all methods show rather similar results, DIFboost compares well to the competitors. This also holds for the settings where no DIF is present. Again, Lord shows the lowest FPR and by far does not reach the intended α -level of 5%.

Setting			DIFboost	MH	Lord	Logistic		
1	$P = 500$ $I = 20$ $I_{\text{DIF}} = 4$ $k = 2$	very strong	TPR FPR	0.725 0.030	0.765 0.037	0.733 0.025	0.810 0.049	
		strong	TPR FPR	0.305 0.041	0.292 0.034	0.260 0.026	0.343 0.048	
		medium	TPR FPR	0.190 0.041	0.168 0.034	0.147 0.026	0.203 0.046	
		weak	TPR FPR	0.117 0.041	0.087 0.037	0.085 0.026	0.140 0.048	
	$I_{\text{DIF}} = 0$	no DIF	FPR	0.041	0.037	0.024	0.445	
	2	$P = 500$ $I = 20$ $I_{\text{DIF}} = 8$ $k = 2$	very strong	TPR FPR	0.705 0.019	0.782 0.044	0.757 0.033	0.823 0.051
strong			TPR FPR	0.281 0.029	0.300 0.034	0.258 0.026	0.347 0.047	
medium			TPR FPR	0.198 0.036	0.179 0.035	0.161 0.027	0.217 0.045	
weak			TPR FPR	0.114 0.040	0.095 0.037	0.080 0.024	0.133 0.042	
3*		$P = 500$ $I = 20$ $I_{\text{DIF}} = 4$ $k = 2$	very strong	TPR FPR	0.677 0.034	0.685 0.044	0.692 0.031	0.735 0.062
			strong	TPR FPR	0.212 0.045	0.195 0.041	0.185 0.031	0.258 0.059
	medium		TPR FPR	0.150 0.048	0.128 0.040	0.120 0.031	0.170 0.059	
	weak		TPR FPR	0.075 0.051	0.082 0.043	0.065 0.029	0.100 0.059	
	$I_{\text{DIF}} = 0$	no DIF	FPR	0.048	0.041	0.029	0.056	
	4	$P = 500$ $I = 20$ $I_{\text{DIF}} = 4$ $k = 5$	strong	TPR FPR	0.990 0.027	1.000 0.049	0.993 0.017	1.000 0.058
medium			TPR FPR	0.875 0.026	0.910 0.051	0.845 0.015	0.927 0.056	
weak			TPR FPR	0.570 0.031	0.593 0.049	0.470 0.016	0.608 0.053	
$I_{\text{DIF}} = 0$			no DIF	FPR	0.047	0.052	0.017	0.51
5		$P = 500$ $I = 20$ $I_{\text{DIF}} = 8$ $k = 5$	strong	TPR FPR	0.976 0.008	0.999 0.072	0.995 0.027	1.000 0.077
			medium	TPR FPR	0.866 0.008	0.944 0.062	0.884 0.020	0.942 0.063
	weak		TPR FPR	0.552 0.012	0.624 0.052	0.471 0.017	0.645 0.055	

Table 5.1.: True positive rates (TPR) and false positive rates (FPR) from five different simulation settings comparing DIFboost to the reference methods MH, Lord and Logistic regression

* the person abilities from Setting 3 are highly correlated with the group membership.

As an additional investigation, we compare the methods by the help of ROC-curves where the TPR is plotted against the FPR, see Magis et al. (2015) for a similar analysis. For that purpose, the settings presented above were used, but with varying parameters. For the reference methods, the level of significance was varied while for DIFboost we varied the threshold parameter π_0 . The goal was to provide a comparison of the methods that

is not confounded by the choice of these parameters. Exemplarily, Figure 5.4 shows the

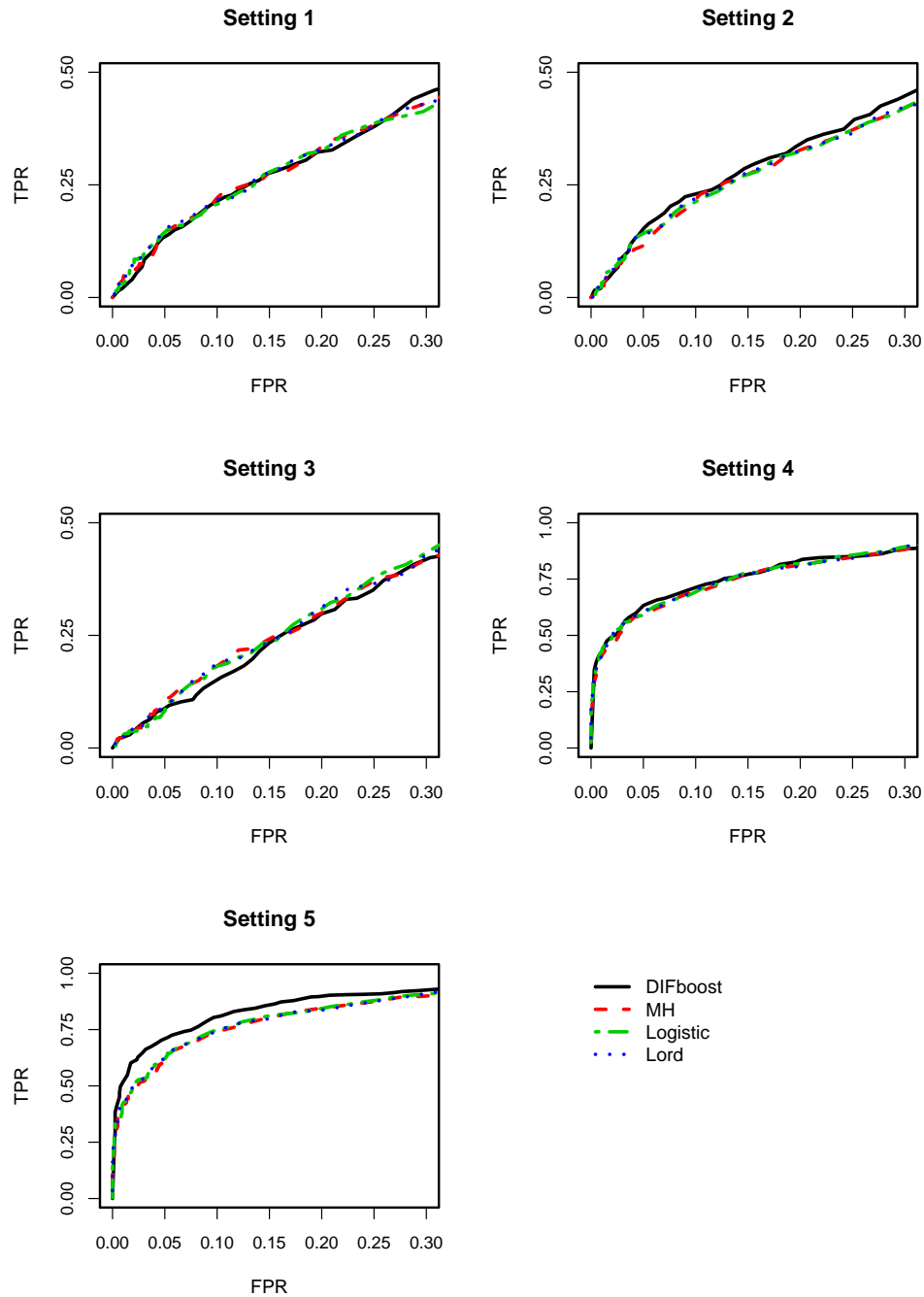


Figure 5.4.: ROC-curves for all weak settings in the simulation study comparing DIFboost to the reference methods MH, Lord and Logistic regression

ROC-curves for all weak settings, the ROC-curves for the other strengths shows similar tendencies and are dropped for the sake of brevity. Again, it can be seen that in general

the performance of all methods is very similar. After all, two tendencies can be seen from the curves. First, DIFboost handles situations with many DIF items better than its competitors. Therefore, it outperforms its competitors in setting 2 and especially in setting 5. Relative to its competitors, it improves from setting 1 to setting 2 and also from setting 4 to setting 5 when 8 instead of 4 items have DIF. Second, DIFboost seems to perform better in more complex situations with more than two groups. Relative to its competitors, it improves from setting 1 to setting 4 and from setting 2 to setting 5 (when $k = 2$ is changed to $k = 5$). Finally, in setting 5 (with both $k = 5$ and $I_{\text{DIF}} = 8$) DIFboost clearly outperforms its competitors.

5.5.2. Simulations with Many Covariates

Methods

As DIFboost can include many covariates at the same time and is able to handle continuous covariates, the method can be used in much more general settings than explored in the previous section. In the following, we present a simulation study for settings, where several possibly DIF-inducing covariates are available. The reference methods from the previous section cannot be used in these situations. Consequently, we compare the methods to the method of DIFlasso from Chapter 4.

The method of Rasch trees (Strobl et al., 2015) can also handle several (possibly continuous) variables simultaneously. After all, this method only provides groups within the respondents with equal item parameters. It does not provide an actual identification of DIF items as, between different groups, all item parameters are different. Therefore, this method can not be used for comparison when it comes to identification of DIF items and will not be used in the simulation study.

Settings

Four different settings are considered, each with $I = 20$ items and $m = 5$ covariates (2 binary, 3 continuous). Again, abilities θ and difficulties β are drawn from standard normal distributions. The number of persons and the number of DIF items are varied. For each setting, ‘strong’, ‘medium’ and ‘weak’ DIF is used with DIF strengths 0.3, 0.15 and 0.1125. In the following, we present the four used settings and the used parameters for ‘medium’ DIF, for a different DIF strength the parameters are simply multiplied by a suited factor.

Setting 1 $P = 250$ persons, $I_{\text{DIF}} = 4$ DIF items, $\gamma_1 = (-0.5, 0.4, 0, 0, 0.5)$,
 $\gamma_2 = (0, 0.5, -0.4, 0, 0.3)$, $\gamma_3 = (0.4, 0, 0.5, -0.5, 0)$, $\gamma_4 = (0, 0, 0.5, 0.4, -0.2)$,
 $\gamma_5, \dots, \gamma_{20} = (0, 0, 0, 0, 0)$

Setting 2 same as **Setting 1**, but with $P = 500$ persons

Setting 3 same as **Setting 2**, but with $I_{\text{DIF}} = 8$ DIF items, items 5–8 same as items 1–4

Setting 4 same as **Setting 2**, but the abilities are highly correlated with the group membership: $\theta_i|x_i = 0 \sim N(0, 1)$, $\theta_i|x_i = 1 \sim N(1, 1)$

Again, for settings 1,2 and 4 also no-DIF settings are simulated, where 2 differs from 4 as in the latter the abilities are correlated with the group membership.

Results

Table 5.2 shows the results for 100 replications of the different simulation settings in terms of true positive rates (TPR) and false positive rates (FPR).

Setting			DIFboost	DIFlasso
1	$P = 250$	strong	TPR	1.000
			FPR	0.024
	$I = 20$	medium	TPR	0.873
			FPR	0.028
	$I_{\text{DIF}} = 4$	weak	TPR	0.642
			FPR	0.029
	$I_{\text{DIF}} = 0$	no DIF	FPR	0.053
2	$P = 500$	strong	TPR	1.000
			FPR	0.011
	$I = 20$	medium	TPR	1.000
			FPR	0.029
	$I_{\text{DIF}} = 4$	weak	TPR	0.948
			FPR	0.026
	$I_{\text{DIF}} = 0$	no DIF	FPR	0.051
3	$P = 500$	strong	TPR	1.000
			FPR	0.002
	$I = 20$	medium	TPR	0.990
			FPR	0.007
	$I_{\text{DIF}} = 8$	weak	TPR	0.900
			FPR	0.008
	$m = 5$			
4*	$P = 500$	strong	TPR	1.000
			FPR	0.016
	$I = 20$	medium	TPR	0.968
			FPR	0.031
	$I_{\text{DIF}} = 4$	weak	TPR	0.873
			FPR	0.033
	$m = 5$			
	$I_{\text{DIF}} = 0$	no DIF	FPR	0.065

Table 5.2.: True positive rates (TPR) and false positive rates (FPR) for four different simulation settings comparing DIFboost to DIFlasso

* the person abilities from Setting 4 are highly correlated with one of the binary covariates.

For medium and especially for weak DIF, DIFboost clearly outperforms DIFlasso in terms of TPR. Also, DIFlasso shows increased FPR in some settings whereas DIFboost is very stable regarding FPR. Therefore, DIFboost proved to be a very interesting alternative regarding the DIF detection for several covariates. For the settings with no DIF, it is no surprise that DIFlasso has a lower FPR than DIFboost. Still, the chosen parameters for DIFboost provide FPRs around 5% and, therefore, if no DIF is present the procedure can be compared to a test procedure with a level of significance $\alpha = 0.05$.

5.6. DIF in the Intelligence-Structure-Test 2000 R

In the following, the method is applied to data from the Intelligence-Structure-Test 2000 R (I-S-T 2000 R; source of supply is Testzentrale Göttingen, Herbert-Quandt-Str. 4, 37081 Göttingen, Tel. (0049-551) 999-50-999, www.testzentrale.de), developed by Amthauer et al. (2001). The test is a fundamentally revised version of its predecessors I-S-T 70 (Amthauer et al., 1973) and I-S-T 2000 (Amthauer et al., 1999). Generally, it aims at measuring the ability of deductive reasoning of the participants. It consists of three basic modules on verbal intelligence, numerical intelligence and figural intelligence. Each of these modules is divided into three subtests where each subtest consists of 20 items. For example, the module for numerical intelligence consists of the subtests numerical calculations, number series and numerical signs. Further details on the I-S-T 2000 R and its predecessors can be found, for example, in Schmidt-Atzert et al. (1995), Brocke et al. (1998) and Schmidt-Atzert (2002).

The data origin from a test on 273 students from different faculties from the university of Marburg, Germany, aged between 18 and 39 years. The data have already been analysed in Bühner et al. (2006), where the data were used to test if the I-S-T 2000 R is Rasch-scalable using mixed Rasch models (Rost, 1990).

We will analyse the items of the subtest sentence completion from the module verbal intelligence. Three covariates were used as possibly DIF inducing covariates, gender (0: male, 1: female), age (in years) and the interaction between gender and age.

Figure 5.5 shows the stability paths for DIFboost, where, in accordance with the simulation study, the parameters $q = 0.6 \cdot I = 12$ and $\pi_0 = 0.9$ are chosen. It is seen, that four items are identified to have DIF, namely the Items 8, 9, 11 and 15.

We illustrate the coefficients of the DIF-items by effect stars, see also Appendix A. Since the logit link is used, the exponentials of the coefficients represent the effects of the covariates on the odds $\frac{P(Y_{pi}=1)}{P(Y_{pi}=0)}$. The length of the rays corresponds to the exponentials of the respective coefficients. The circle around each star has a radius of $\exp(0) = 1$ and, therefore, represents the no-effect case. Both gender and age were standardized prior to the analysis so that the

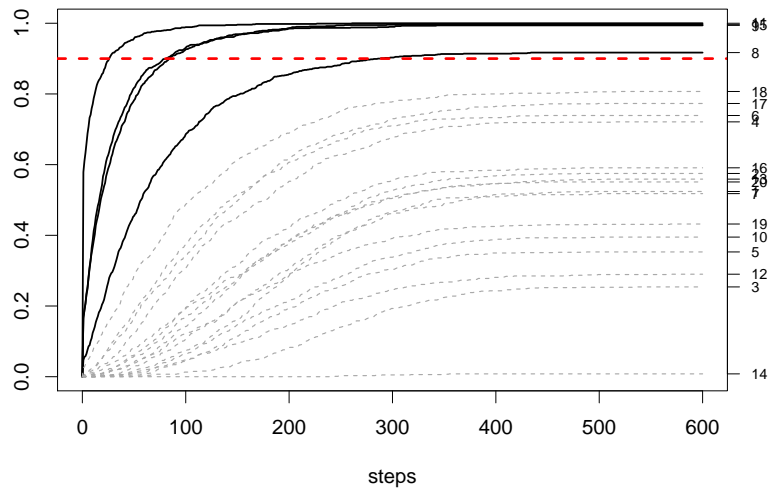
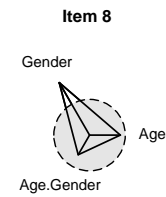


Figure 5.5.: Stability paths for DIFboost for the items of the subtest sentence completion; dashed line represents the threshold $\pi_0 = 0.9$; items 8, 9, 11 and 15 are diagnosed as DIF items

size of the coefficient estimates is comparable. Figure 5.6 shows the effect stars for the estimated coefficients and the item descriptions of the DIF items.

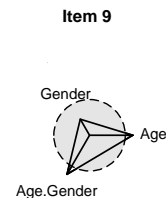
Generally, a ray beyond the circle represents positive coefficients. With positive coefficients, the difficulty of the respective item is increased if the corresponding covariate is increased while the probability to solve the item is decreased. Item 9, for example, has a negative coefficient for gender. Therefore, this item is easier for female participants as female is encoded by 1. After all, since also the interaction between gender and age is considered, one has to look at all coefficients at a time. With growing age, the difficulty increases for female participants.

Figure 5.7 shows for each DIF-item the effects of both gender and age on the probability to score on the respective item. Separately for male (solid lines) and female (dashed lines) participants, the probability to score on the respective item is depicted along the covariate age. For simplicity, the plots refer to a person with a 'mean' ability according to the estimates of the θ parameters. Figure 5.7 demonstrates the effect of the interaction term. As the probabilities to score on an item can intersect, the main effects of age or gender should not be interpreted separately but always with respect to the interaction term. The ability to include interaction terms in this manner can be seen as a big improvement compared to existing methods of DIF detection allowing for new insights on the occurrence of DIF. In extreme cases, both the main effects for gender and age could even be negligible but the interactions term could still be influential.



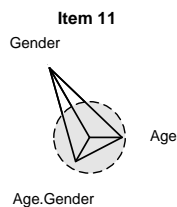
Mercury is a/an ...?

- a) metal b) mineral c) solution d) mixture e) alloy



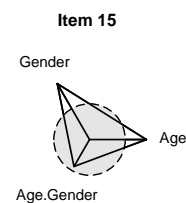
Fathers are ...? (more) experienced than their sons.

- a) always b) usually c) much d) less e) fundamentally



Every river has ...?

- a) fishes b) bridges c) ships d) gradients e) rapids



A watch always needs (a) ...?

- a) battery b) case c) numbers d) energy e) hands

Figure 5.6.: Effect stars and item descriptions for items with DIF in the subtest sentence completion (IST 2000 R, Amthauer et al., 2001) detected by DIFboost

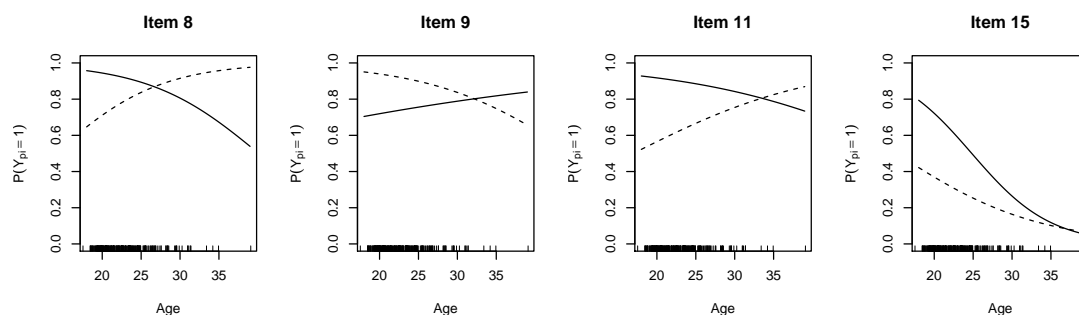


Figure 5.7.: Probabilities to score on items depending on gender and age for all DIF-items. Solid lines represent male, dashed lines represent female participants.

Therefore, item 9 can not generally be assumed to be easier for female participants. This holds only for participants younger than 30 years, but the order changes for older participants. Items 11 and 15 are, in general, easier for male participants, in particular if they

are rather young. For growing age this difference slowly vanishes, in item 11 the effect is even reversed for higher age.

For comparison, the data also were analyzed with the method of Rasch trees (Strobl et al., 2015), the corresponding Rasch tree is plotted in Figure 5.8.

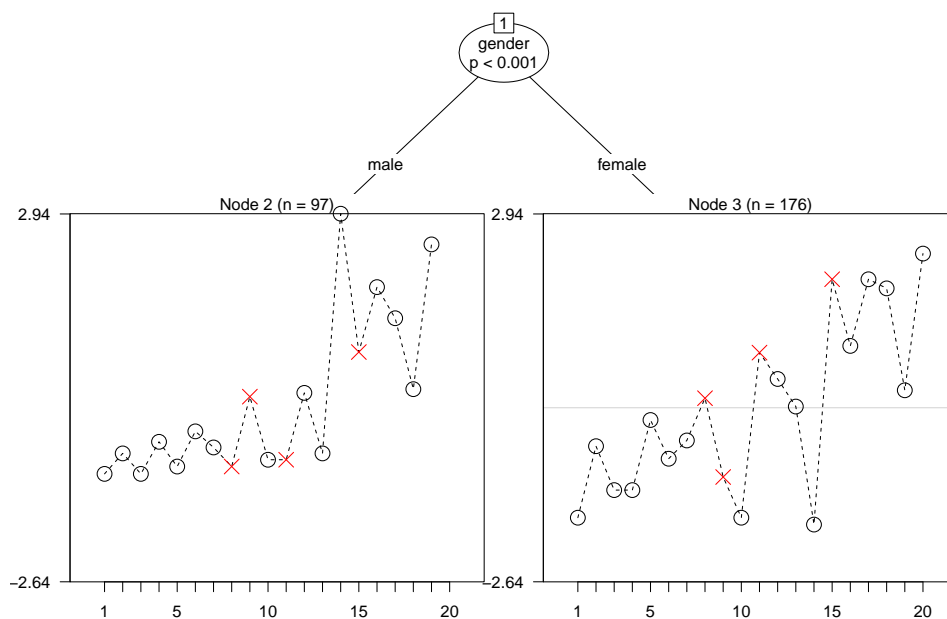


Figure 5.8.: Rasch tree for subtest sentence completion. Highlighted items represent items diagnosed to be DIF items by other methods.

By recursive partitioning of the covariate space, one tries to find groups within the observations which have the same item parameters. In our case, only one partition was found in the data, namely male and female participants. For age, no significant difference was found. When using recursive partitioning, the item parameters are estimated separately within the groups. The estimates are also shown in Figure 5.8. The estimates for the items diagnosed as DIF items from the other methods are highlighted. But, by far the highest difference between both groups seems to be for item 14 which is the hardest item for male participants and the easiest item for female participants. However, all other methods did not identify item 14 to be a DIF item. That means, Rasch trees may yield quite different results than other methods when trying to identify DIF items. To our knowledge, no systematic investigation that compares Rasch trees and alternative methods is available.

5.7. Concluding Remarks

A new method called DIFboost is proposed to detect DIF that is induced by several covariates simultaneously. In the case of DIF in subgroups, the method competes well with established methods for DIF detection. For the more general case of several, possibly continuous covariates, it outperforms the competitive DIFlasso approach from Chapter 4.

In contrast to the established test procedures, DIFboost is able to identify DIF items without the specification of anchor items. In most other methods, one assumes that the other items have no DIF and, therefore, all items besides the investigated item serve as anchor items. Besides this strategy, there exist other possibilities to find anchor items, see, e.g., Kopf et al. (2015) or Woods (2009). After all, the need for anchor items remains problematic, especially if many possible covariates have to be considered.

DIFboost is a model-based method. This provides two further advantages over test-based methods. First, the problem of multiple testing is avoided. Generally, DIF tests perform one test per item and covariate. A test is designed to restrict the probability of a type I error to a certain level. After all, if there are many covariates and many items, there are many tests and the problem of multiple testing arises. To control for that, correction strategies as, for example, the Bonferroni adjustment become necessary.

Second, unlike tests, the DIFboost method provides parameter estimates which allow for a deeper look into the data structure and gives interpretable results. The linear effects of the covariates can be complemented by the incorporation of interaction effects or by using smooth functions for the covariates effects. Therefore, model-based methods do not only tell us which items have DIF but also provide valuable information about the underlying covariate effects.

For simplicity, the presented approach is limited to the Rasch model. However, extensions to other models are possible and should be investigated in future research. For example, the boosting algorithm seems well suited to an extension to the 2PL model. The parameter estimation in the 2PL model is rather complicated because of its multiplicative structure. By using boosting concepts, this problem can be tackled in a stepwise way. In particular, the discrimination parameters can be used as further base learners that are updated only for those items that call for it.

6. The Bradley–Terry Model

In the previous chapters, concepts for the inclusion of covariates into item response models, in particular into the Rasch model, were treated. The proposed methods allow for a generalization of the Rasch model by including additional information of the subjects into the model. In the following chapters, similar concepts for paired comparison models will be proposed. Therefore, in this chapter the Bradley-Terry model as the most popular model for paired comparison data will be introduced.

6.1. The Basic Bradley–Terry Model

The Bradley-Terry model is the indisputable standard model for the modeling of paired comparison data, see Agresti (2002) and Bradley (1984) for basic introductions. Originally, it was proposed by Bradley and Terry (1952). Sometimes, the Bradley-Terry model is also referred to as the Bradley-Terry-Luce (BTL) model indicating the connection of the model to Luce's choice axiom formulated in Luce (1959). Luce's choice axiom states that the decision between two objects is not influenced by other objects. This statement is also known as the independence from irrelevant alternatives.

As indicated in Chapter 1, the Bradley-Terry model has a strong connection to the Rasch model (see Chapter 2) and can be seen as its counterpart for homogeneous (instead of heterogeneous) paired comparisons. Assuming a set of items $\{a_1, \dots, a_m\}$, in its most simple form the Bradley-Terry model is denoted by

$$P(a_r \succ a_s) = P(Y_{(r,s)} = 1) = \frac{\exp(\gamma_r - \gamma_s)}{1 + \exp(\gamma_r - \gamma_s)}.$$

The response of the model represents the probability that a certain item a_r is preferred over another item a_s , $a_r \succ a_s$. This response can be formalized in the random variable $Y_{(r,s)}$ which is defined to be $Y_{(r,s)} = 1$ if a_r is preferred over a_s and $Y_{(r,s)} = 0$ otherwise. The parameters γ_r , $r = 1, \dots, m$, represent the attractiveness or strength of the respective items. For identifiability, a restriction on the parameters is needed, for example $\sum_{r=1}^m \gamma_r = 0$ or $\gamma_m = 0$.

The basic Bradley-Terry model can be embedded into the framework of generalized linear models (GLMs) and is simply estimated as a binary logit model. The linear predictor η_{rs} can simply be rewritten using dummy variables for the objects involved in the respective pair (r, s) by

$$\eta_{(rs)} = \gamma_r - \gamma_s = x_1^{(r,s)}\gamma_1 + \cdots + x_m^{(r,s)}\gamma_m = (\mathbf{x}^{(r,s)})^T \boldsymbol{\gamma}.$$

Here, the components of the vector $\mathbf{x}^{(r,s)} = (x_1^{(r,s)}, \dots, x_m^{(r,s)})$ are given by

$$x_j^{(r,s)} = \begin{cases} 1 & j = r \\ -1 & j = s \\ 0 & \text{otherwise.} \end{cases}.$$

Using these variables, an appropriate design matrix can be built and standard software for generalized linear models, more precisely for binomial logit models, can be used for estimation.

6.2. Extensions of the Bradley–Terry Model

Tutz (1986) and Agresti (1992) extended the Bradley-Terry model to the case of ordered response, for example to allow for a 5-point scale (much better, slightly better, equal, slightly worse, much worse). In particular, a category for ties is often necessary for various applications, for example in sport competitions. For K different response categories, the model considers the cumulative probabilities

$$P(Y_{(r,s)} \leq k) = \frac{\exp(\theta_k + \gamma_r - \gamma_s)}{1 + \exp(\theta_k + \gamma_r - \gamma_s)}$$

with $k = 1, \dots, K$ denoting the possible response categories. The parameters θ_k represent the so-called threshold parameters for the single response categories, they determine the preference for specific categories. In particular, $Y_{(r,s)} = 1$ represents the maximal preference for item a_r over a_s and $Y_{(r,s)} = K$ represents the maximal preference for item a_s over a_r . In general, for ordinal paired comparisons it can be assumed that the response categories have a symmetric interpretation so that $P(Y_{(r,s)} = k) = P(Y_{(s,r)} = K - k + 1)$ holds. Therefore, the threshold parameters should be restricted with $\theta_k = -\theta_{K-k}$ and, if K is even, $\theta_{K/2} = 0$ to guarantee for symmetric probabilities. The threshold for the last category is fixed to $\theta_K = \infty$ so that $P(Y_{(r,s)} \leq K) = 1$ will hold. The probability for a single response category can be derived from the difference between two adjacent categories, $P(Y_{(r,s)} = k) = P(Y_{(r,s)} \leq k) - P(Y_{(r,s)} \leq k - 1)$. To guarantee for non-negative probabilities for the single response categories one restricts $\theta_1 \leq \theta_2 \leq \dots \leq \theta_K$. The ordinal

Bradley-Terry model corresponds to a cumulative logit model and can be estimated using methods from this general framework.

In some specific paired comparisons it can be decisive in which order the competing items are presented. Typical examples are sports events. Here, the first-mentioned team typically is the team playing at its home ground where it might have a (home) advantage over its opponent. Therefore, the assumption that the response categories are symmetric does not hold anymore and the model needs to be adapted accordingly. Extending the basic models by an additional parameter δ , the binary Bradley-Terry model is denoted by

$$P(Y_{(r,s)} = 1) = \frac{\exp(\delta + \gamma_r - \gamma_s)}{1 + \exp(\delta + \gamma_r - \gamma_s)}$$

and the ordinal model is denoted by

$$P(Y_{(r,s)} \leq k) = \frac{\exp(\delta + \theta_k + \gamma_r - \gamma_s)}{1 + \exp(\delta + \theta_k + \gamma_r - \gamma_s)}.$$

Here, δ denotes the order effect which is simply incorporated into the design matrix by an additional intercept column. If $\delta > 0$, it increases the probability of the first-named object a_r to win the comparison or, in case of an ordinal response, to achieve a good result. Given the order effect, the symmetry assumption for the response categories still holds.

7. Modelling Heterogeneity in Paired Comparison Data

7.1. Introduction

Paired comparisons are a well established method to measure the relative preference or dominance of objects or items. The aim is to find the underlying preference scale by presenting the items in pairs. The method has been used in various areas, for example, in psychology, to measure the intensity or attractiveness of stimuli, in marketing, to evaluate the attractiveness of brands, in social sciences, to investigate the value orientation (e.g. Francis et al. (2002)). In all these applications the items or stimuli are presented in an experiment. But paired comparison are also found in sports whenever two players or teams compete in a tournament. Then, the non-observable scale to be found refers to the strengths of the competitors. Paired comparisons can be obtained from ranked data (Francis et al., 2010) or from scale data (Dittrich et al., 2007). In this kind of data, respondents rank a predefined number of items or assign values from a Likert scale to the items, always referring to a certain attitude of the respondents towards the items. Building differences between the ranks or scales yields (binary or ordered) paired comparison data. We consider an application that shows how to analyse scales for the preference of parties by paired comparisons. In a German pre-election study the respondents were asked to scale the most renowned German parties. The focus of the analysis is on the inclusion of subject-specific covariates to account for the heterogeneity in the population and to investigate which variables determine the preference. More precisely, we investigate which clusters of parties are distinguished by specific covariates allowing that some covariates have no effect on the preference at all.

The most widely used model for paired comparison data is the Bradley-Terry-Luce model. It has been proposed by Bradley and Terry (1952) and is strongly linked to Luce's choice axiom (Luce, 1959). The basic model has been extended in various ways allowing for dependencies among responses, time dependence or simultaneous ranking with respect to more than

This chapter is a modified version of the technical report 183 (Schauberger and Tutz, 2015c), previous work on the issue can be found in the conference paper Schauburger and Tutz (2015a). See Chapter 1 for more information on the personal contributions of all authors and textual matches.

one attribute. Overviews are found in the review of Bradley (1976), the monograph of David (1988) and more recently in the review of Cattelan (2012). The method proposed in this chapter can be applied both to binary and ordered response. Former approaches for ordered responses in paired comparisons include Tutz (1986) and Agresti (1992). Dittrich et al. (2004) also combine ordered responses and the inclusion of covariates, yet in a quite different modelling approach using log-linear models and without variable selection.

When persons choose between a pair of items most models assume that the strengths of the items are fixed and equal for all persons. Heterogeneity over persons has rarely been modeled explicitly. Exceptions are Turner and Firth (2012) or Francis et al. (2010), where categorical covariates are considered, but the application is very low dimensional with just two covariates, one with two and one with four categories. Also in Francis et al. (2002) covariates are included. Their model allows even for smooth effects of subject-specific covariates, but the fitting procedure that is proposed is also restricted to few variables. More recently, Casalicchio et al. (2015) presented a boosting approach that is able to include explanatory variables. An alternative approach has been proposed by Strobl et al. (2011). It is based on recursive partitioning techniques (also known as trees) and automatically selects the relevant variables among a potentially large set of variables.

The method proposed here is an alternative to handle the inherently high dimensional estimation problem that comes with the inclusion of explanatory variables. Maximum likelihood estimation is replaced by penalized estimation methods. By using a specific L_1 -type penalty, the method is able to fit in high dimensional settings and to form clusters of items regarding the variables that generate heterogeneity.

In Section 7.2 the basic Bradley-Terry-Luce model for binary and ordered response is introduced. Then the model is extended to include subject-specific covariates. Section 7.3 contains the integration of the proposed model into the framework of generalized linear models and the penalty term is introduced. Section 7.3 also describes the implementation of the algorithm, the search for the optimal tuning parameter and the calculation of bootstrap confidence intervals. In Section 7.4, the method is applied to data from the German Longitudinal Election Study (GLES).

7.2. Bradley-Terry Models with Covariates

7.2.1. The Basic Model

Let $\{a_1, \dots, a_m\}$ denote the set of objects or items to be compared in a paired comparison experiment. The basic Bradley-Terry model (Bradley and Terry, 1952) specifies the probability that item a_r is preferred over a_s as

$$P(a_r \succ a_s) = \frac{\exp(\gamma_r - \gamma_s)}{1 + \exp(\gamma_r - \gamma_s)},$$

where, for reasons of identifiability, we use the restriction $\sum_{r=1}^m \gamma_r = 0$. The parameters γ_r , $r = 1, \dots, m$, represent the attractiveness of the items $\{a_1, \dots, a_m\}$. The interpretation as strength parameters is straightforward. For $\gamma_r = \gamma_s$, the probability that a_r is preferred over a_s is 0.5, for growing distance $\gamma_r - \gamma_s$ the probability increases.

With the random variable $Y_{(r,s)} = 1$ if $r \succ s$ and $Y_{(r,s)} = 0$ otherwise one obtains the logit model

$$\log \left(\frac{P(Y_{(r,s)} = 1)}{P(Y_{(r,s)} = 0)} \right) = \gamma_r - \gamma_s.$$

7.2.2. Bradley-Terry Models with Ordered Response

In some applications, paired comparison data can or should not be reduced to binary decisions. For example in sport events like football matches where also draws are possible, simple binary paired comparisons are not appropriate. A model that allows for ordinal responses is the cumulative Bradley-Terry-Luce model (Tutz, 1986) which has the form

$$P(Y_{(r,s)} \leq k) = \frac{\exp(\theta_k + \gamma_r - \gamma_s)}{1 + \exp(\theta_k + \gamma_r - \gamma_s)} \quad (7.1)$$

with the same restriction $\sum_{r=1}^m \gamma_r = 0$.

The parameters $\theta_1, \dots, \theta_K$ represent threshold parameters for the different levels of the response $Y_{(r,s)} \in \{1, \dots, K\}$. The response $Y_{(r,s)} = 1$ corresponds to a strong preference of a_r over a_s and $Y_{(r,s)} = K$ corresponds to a strong preference of a_s over a_r . The basic Bradley-Terry model can be seen as a special case of model (7.1) for binary response with $K = 2$.

The strength parameters $\gamma_1, \dots, \gamma_m$ have the same interpretation as in the binary model. With increasing γ_r the probability for low response categories, and therefore the strong preference of a_r over a_s is increasing while the probability for large response categories denoting dominance of a_s decreases. The threshold parameters determine the preference for specific categories. The threshold for the last category K is restricted to $\theta_K = \infty$ so that $P(Y_{(r,s)} \leq K) = 1$ holds. It is sensible to put further restrictions on the threshold parameters to ensure equal probabilities for corresponding categories if the order of the paired comparison is reversed. Therefore, we use the restrictions $\theta_k = -\theta_{K-k}$ and, if K is even, additionally $\theta_{K/2} = 0$. These restrictions ensure, for example, that $Y_{(r,s)} = 1$ (maximal preference of a_r over a_s) has the same probability as $Y_{(s,r)} = K$. Due to these restrictions, $\lfloor \frac{K-1}{2} \rfloor$ (free) threshold parameters have to be estimated. In the special case

of binary response ($K = 2$) all threshold parameters are omitted and the model reduces to the ordinary Bradley-Terry model. If an order effect is required, for example to model the home advantage in sport competitions, an additional parameter can be included. For the application considered here no order effect is needed and therefore is omitted.

Formally, model (7.1) is a cumulative logit model, also called a proportional odds model. For a response variable consisting of K ordered categories, one models $K - 1$ cumulative probabilities $P(Y_{(r,s)} \leq 1), \dots, P(Y_{(r,s)} \leq K - 1)$. The probability for a single response category is represented by the difference $P(Y_{(r,s)} = k) = P(Y_{(r,s)} \leq k) - P(Y_{(r,s)} \leq k - 1)$. Therefore, $P(Y_{(r,s)} \leq k)$ has to be greater or equal $P(Y_{(r,s)} \leq k - 1)$ for $k = 1, \dots, K$ to have non-negative probabilities for all single categories. As the probabilities only differ with respect to the threshold parameters, this is ensured if $\theta_1 \leq \theta_2 \leq \dots \leq \theta_K$.

7.2.3. Heterogeneity in the Bradley-Terry Model

The models considered so far assume that all persons have the same preference structure. Heterogeneity in the population is simply ignored. A more sensible assumption is that preferences depend on covariates that characterize the person that chooses.

Let $Y_{i(r,s)}$ denote the response of person i for given pair of items (r, s) and $\mathbf{x}_i^T = (x_{i1}, \dots, x_{ip})$ be a person-specific covariate vector. It is assumed that the strength of the preference of item a_r for person i is determined by $\gamma_{ir} = \beta_{r0} + \mathbf{x}_i^T \boldsymbol{\beta}_r$. That means there is a global strength parameter β_{r0} but the effective strength is modified by the covariates. The parameter $\boldsymbol{\beta}_r^T = (\beta_{r1}, \dots, \beta_{rp})$ contains the effect of the covariates on item a_r . The corresponding model has the form

$$\begin{aligned}
 P(Y_{i(r,s)} \leq k \mid \mathbf{x}_i) &= \frac{\exp(\theta_k + \gamma_{ir} - \gamma_{is})}{1 + \exp(\theta_k + \gamma_{ir} - \gamma_{is})} \\
 &= \frac{\exp(\theta_k + (\beta_{r0} + \mathbf{x}_i^T \boldsymbol{\beta}_r) - (\beta_{s0} + \mathbf{x}_i^T \boldsymbol{\beta}_s))}{1 + \exp(\theta_k + (\beta_{r0} + \mathbf{x}_i^T \boldsymbol{\beta}_r) - (\beta_{s0} + \mathbf{x}_i^T \boldsymbol{\beta}_s))} \\
 &= \frac{\exp(\theta_k + \beta_{r0} - \beta_{s0} + \mathbf{x}_i^T (\boldsymbol{\beta}_r - \boldsymbol{\beta}_s))}{1 + \exp(\theta_k + \beta_{r0} - \beta_{s0} + \mathbf{x}_i^T (\boldsymbol{\beta}_r - \boldsymbol{\beta}_s))} \quad (7.2)
 \end{aligned}$$

As in model (7.1), the sum-to-zero constraints $\sum_{r=1}^m \beta_{rj} = 0$ with $j = 0, 1, \dots, p$ are used for identifiability.

The model allows for different preference structures in sub populations. For illustration let us consider the simple case where the person-specific variable codes a subgroup like gender,

which has two possible values. Let $x_i = 1$ for males and $x_i = 0$ for females. Then the strengths parameters for item r are

$$\beta_{r0} + \beta_r \text{ for males and } \beta_{r0} \text{ for females.}$$

The β_r represents the difference in attractiveness of item a_r between males and females. When items a_r and a_s are compared the dominance in the male population is determined by $(\beta_{r0} - \beta_{s0}) + (\beta_r - \beta_s)$, in the female population by $(\beta_{r0} - \beta_{s0})$. Thus the female population is like a reference population with dominance determined by the difference in the basic parameters $(\beta_{r0} - \beta_{s0})$. The preference in the male population is modified by the term $\beta_r - \beta_s$, and can be quite different. If one prefers a more symmetric representation one can choose $x_i = 1$ for males and $x_i = -1$ for females obtaining for the strengths parameters for item r

$$\beta_{r0} + \beta_r \text{ for males and } \beta_{r0} - \beta_r \text{ for females.}$$

Then β_r represents the deviation of the attractiveness of item r from the baseline attractiveness β_{r0} . When items a_r and a_s are compared the dominance in the male population is determined by $(\beta_{r0} - \beta_{s0}) + (\beta_r - \beta_s)$, in the female population by $(\beta_{r0} - \beta_{s0}) - (\beta_r - \beta_s)$. Thus the difference of the basic parameters $\beta_{r0} - \beta_{s0}$ is augmented by $\beta_r - \beta_s$ in the male population and reduced by the same value in the female population.

The model accounts for the heterogeneity in the population by explicitly linking the attractiveness of alternatives to explanatory variables. The weight parameters β_r reflect how the attractiveness of a specific alternative depends on the covariates.

7.3. Penalized Estimation

The main problem with the general model (7.2) is the number of parameters that are involved. One has (with the given restrictions) $\left\lfloor \frac{K-1}{2} \right\rfloor$ threshold parameters and for each item the $(p+1)$ -dimensional parameter vector (β_{r0}, β_r) . In general, not all covariates might have a (different) influence on all m items. Therefore, we propose to use a penalized likelihood approach instead of ordinary maximum likelihood estimation to reduce the number of involved parameters and to select the relevant variables. In a first step we embed the estimation into the framework of generalized linear models (GLMs) and then introduce penalty terms.

7.3.1. Embedding into Generalized Linear Models

First, the ordinal Bradley-Terry model is embedded into the framework of Generalized Linear Models (GLMs). In the ordinal Bradley-Terry model without covariates the linear predictor $\eta_{(r,s)k} = \theta_k + \gamma_r - \gamma_s$ can be given as

$$\eta_{(r,s)k} = \theta_k + x_1^{(r,s)}\gamma_1 + \cdots + x_m^{(r,s)}\gamma_m = \theta_k + (\mathbf{x}^{(r,s)})^T \boldsymbol{\gamma},$$

where $x_l^{(r,s)} = 1$ if $l = r$, $x_l^{(r,s)} = -1$ if $l = s$, and $x_l^{(r,s)} = 0$ otherwise, encodes the considered pair. The whole vector $\mathbf{x}^{(r,s)}$ has the simple form $\mathbf{x}^{(r,s)} = \mathbf{1}_r - \mathbf{1}_s$, where $\mathbf{1}_r = (0, \dots, 0, 1, 0, \dots, 0)$ has length m with 1 at position r . In this model the strength of an item is the same for all persons, which is a strong assumption ignoring potential heterogeneity.

In the general model with covariates, and therefore explicit modelling of heterogeneity, the linear predictor has the form

$$\begin{aligned} \eta_{i(r,s)k} &= \theta_k + \beta_{r0} - \beta_{s0} + \mathbf{x}_i^T (\boldsymbol{\beta}_r - \boldsymbol{\beta}_s) \\ &= \theta_k + \sum_{j=0}^p x_{ij}(\beta_{rj} - \beta_{sj}) = \theta_k + \sum_{j=0}^p \sum_{l=1}^m x_{ij} x_l^{(r,s)} \beta_{lj} \end{aligned}$$

where $x_{i0} = 1$ is a fixed intercept. Here, $\mathbf{x}_i^T = (x_{i1}, \dots, x_{ip})$ represents a covariate vector associated to person i and, therefore, the linear predictors for the same pair are different for persons. For $j > 0$ the predictor is determined by interactions between x_{ij} and the items, which reflects the underlying structure that the item strength is modified by the covariates.

The link between the linear predictor and the probability $P(Y_{i(r,s)} \leq k \mid \mathbf{x}_i)$ is determined by the logistic distribution function. It should be noted that the ordered response is transformed into a multivariate response $\mathbf{y}_{i(r,s)}^T = (y_{i(r,s)1}, \dots, y_{i(r,s)q})$ with $q = K - 1$ binary variables where $y_{i(r,s)k} = 1$ if $Y_{i(r,s)} \leq k$ and $y_{i(r,s),k} = 0$ if $Y_{i(r,s)} > k$. With $\pi_{i(r,s)k} = \exp(\eta_{i(r,s)k}) / (1 + \exp(\eta_{i(r,s)k}))$, the covariance structure for such a multivariate response is given by

$$Cov(\mathbf{y}_{i(r,s)}) = \begin{pmatrix} \pi_{i(r,s)1}(1 - \pi_{i(r,s)1}) & \pi_{i(r,s)1}(1 - \pi_{i(r,s)2}) & \cdots & \pi_{i(r,s)1}(1 - \pi_{i(r,s)q}) \\ \pi_{i(r,s)1}(1 - \pi_{i(r,s)2}) & \pi_{i(r,s)2}(1 - \pi_{i(r,s)2}) & & \vdots \\ \vdots & & \ddots & \vdots \\ \pi_{i(r,s)1}(1 - \pi_{i(r,s)q}) & \cdots & \cdots & \pi_{i(r,s)q}(1 - \pi_{i(r,s)q}) \end{pmatrix}$$

Because of the restrictions $\theta_k = -\theta_{K-k}$ and, if K is even, $\theta_{K/2} = 0$, the design matrix for the threshold parameters has a special form. As stated above, for a response with K categories, $\lfloor \frac{K-1}{2} \rfloor$ different threshold parameters have to be estimated. Therefore, the part of the design matrix corresponding to the paired comparison (r, s) of one person is a $(K-1) \times \lfloor \frac{K-1}{2} \rfloor$ matrix. This matrix is given by

$$\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & 1 \\ 0 & \cdots & 0 & -1 \\ \vdots & & \ddots & 0 \\ 0 & -1 & & \vdots \\ -1 & 0 & \cdots & 0 \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & 1 \\ 0 & \cdots & \cdots & 0 \\ 0 & \cdots & 0 & -1 \\ \vdots & & \ddots & 0 \\ 0 & -1 & & \vdots \\ -1 & 0 & \cdots & 0 \end{pmatrix}$$

for K uneven or even, respectively. As stated above, for $K = 2$ the model reduces to a GLM with binomial distributed response and all threshold parameters are eliminated from the model.

7.3.2. Selection by Penalization

In regression models with β as the parameter vector penalization approaches maximize the penalized likelihood

$$l_p(\beta) = l(\beta) - \lambda J(\beta),$$

where $l(\beta)$ is the usual log-likelihood and $J(\beta)$ is a penalty term that penalizes specific structures in the parameter vector. The parameter λ is a tuning parameter that specifies how seriously the penalty term has to be taken. A simple penalty term that could be used is the squared length of the parameter vector $J(\beta) = \beta^T \beta = \sum \beta_i^2$, known as ridge penalty, see, for example Hoerl and Kennard (1970), Nyquist (1991), Segerstedt (1992), LeCessie (1992). Then, for $\lambda = 0$ maximization yields the ML estimate. If $\lambda > 0$ one obtains parameters that are shrunk toward zero. For appropriately chosen λ the ridge estimator stabilizes estimates. A disadvantage of the ridge estimator is that it does not select variables. Thus no reduction of the model is obtained. An alternative penalty is the L_1 -penalty, also known as lasso (Tibshirani, 1996), which is able to select variables. Instead of the squared parameters one penalizes the absolute values of the parameters with the penalty term $J(\beta) = \sum |\beta_i|$. For penalized likelihood estimation, it is essential that

all covariates are on comparable scales. Therefore, in the following it is assumed that all covariates are standardized.

However, the simple lasso cannot be used directly since penalty terms for paired comparison models have to account for the specific structure of the model. In particular, in model (7.2) one has the parameters of the regular (ordinal) BTL model, namely the threshold parameters and, for each item r , a parameter β_{r0} for its basic attractiveness. They form the basic model and, therefore, will not be penalized. In the general model one has additional parameters for the interaction between the items and the covariates. These parameters will be penalized to obtain the interactions that are actually needed. The proposed penalty term has the form

$$J(\boldsymbol{\alpha}) = \sum_{j=1}^p \sum_{r < s} w_{rsj} |\beta_{rj} - \beta_{sj}|,$$

where $r, s \in \{1, \dots, m\}$, w_{rsj} is a weight parameter and the parameters are collected in $\boldsymbol{\alpha}^T = (\theta_1, \dots, \theta_{K-1}, \beta_{10}, \dots, \beta_{mp})$. The penalty has the effect that the parameters referring to the same covariate are shrunk towards each other. For large values of λ , the differences are shrunk to exactly zero so that the effect of a covariate is the same for two (or more items). Therefore, the penalty yields clusters of items which share the same effect of a certain covariate. With growing tuning parameter, these clusters become bigger until all items form one single cluster. In that case, due to the sum-to-zero constraints all parameters are zero and the covariate is irrelevant for the attractiveness of the items. The penalty is a L_1 -type fusion penalty rather than a simple lasso. Similar penalties have been used for the modelling of factors in GLMs by Bondell and Reich (2009), Gertheiss and Tutz (2010) and Oelker et al. (2014). More recently, penalties of this form have also been used in the modelling of paired comparison models by Masarotto and Varin (2012) and will also be used in Chapter 8. However, these applications do not use the penalty term for the modelling of heterogeneity by inclusion of covariates.

For illustration, Figure 7.1 shows the coefficient paths corresponding to a covariate j for a toy example with $m = 5$ items. The paths are drawn along the (normed) penalty term $\sum_{r < s} |\beta_{rj} - \beta_{sj}|$ for covariate j . It can be seen that the penalty enforces a clustering of the items when the penalty is increased. In the unpenalized model, all items form clusters of their own. With increasing penalty, items 1 and 4 form a cluster, later item 3 is integrated into that cluster. Next, also items 2 and 5 form a cluster and finally all items form one single cluster. If all items share the same parameter (all parameters are zero) that means that the respective covariate is eliminated from the model. Therefore, the proposed penalty term enforces both clustering of items and variable selection at the same time.

Zou (2006) proposed the so-called adaptive lasso as an extension of the regular lasso. In contrast to regular lasso, it provides consistency in terms of variable selection. In the adaptive lasso, the single penalty terms are weighted with the inverses of the unpenalized

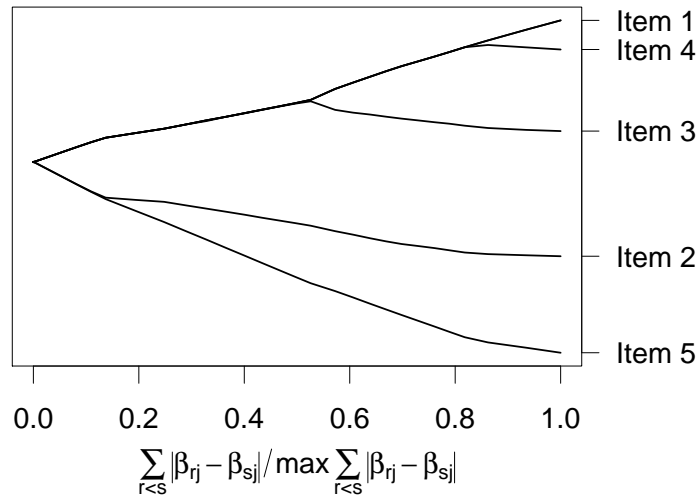


Figure 7.1.: Exemplary coefficient paths for a covariate j in a setting with $m = 5$ different items

ML-estimates. In a similar way the weight parameters w_{rsj} are defined by $w_{rsj} = |\beta_{rj}^{\text{ML}} - \beta_{sj}^{\text{ML}}|^{-1}$. The effect is that small differences in the ML-estimates are penalized stronger than bigger differences which has the effect that the clustering of the parameters is enforced.

7.3.3. Implementation

L_1 penalized cumulative logit models have, e.g., been used in Archer and Williams (2012) and are implemented for R (R Core Team, 2015) in Archer (2014a) and Archer (2014b). However, these implementations are limited to lasso type penalties for coefficients. They cannot be used to penalize differences between parameters as required in the paired comparison case. Moreover, in order to obtain consistent estimates we want to include the weights w_{rsj} . For that purpose, a new fitting algorithm was implemented that is able to fulfill these requirements. It is based on the idea of approximating penalties proposed by Oelker and Tutz (2015), which is implemented in the R-package `gvcm.cat` (Oelker, 2015), yet not for cumulative logit models. For shorter computation time, the fitting algorithm itself is implemented in C++ and integrated into R using the packages `Rcpp` (Eddelbuettel et al., 2011; Eddelbuettel, 2013) and `RcppArmadillo` (Eddelbuettel and Sanderson, 2014). The code is available on CRAN in the R-package `BTLLasso` (Schauberger, 2015a).

7.3.4. Choice of Penalty Parameter

The performance of penalized estimation methods is essentially determined by the choice of the tuning parameter λ . It determines which covariates modify the attractiveness and forms the clusters within the chosen covariates. Mostly, two different approaches are used to determine tunings parameters, namely model selection criteria and cross-validation. Model selection criteria like the AIC (Akaike, 1973) or the BIC (Schwarz, 1978) try to find a compromise between the complexity of the model and the model fit. The complexity of a model is determined by its degrees of freedom. While for ML estimation, the degrees of freedom simply correspond to the number of parameters, the degrees of freedom for penalized likelihood approaches, in particular with a penalty applied on differences, are not straightforward. Therefore, we use cross-validation. In cross-validation, the data set is divided into a predefined number of subsets. Each subset is once used as a test data set while the remaining subsets serve as training data. The model is fitted (for a predefined grid of values for the tuning parameter λ) on the training data while the test data are used for prediction. Then, the predictive performance in the test data can be measured, for example by using the deviance. Moreover, this procedure provides a measure of the predictive performance of the model for every value from the predefined grid of tuning parameters. The tuning parameter with the best performance is chosen. We adapted this general principle to our specific case. The persons or subjects are treated as the observation units so that all paired comparisons corresponding to one person are in the same subset.

7.3.5. Confidence Intervals

In contrast to maximum likelihood estimators, for estimators from penalized likelihood approaches one cannot use the information matrix to obtain standard errors or confidence intervals. Therefore, alternative techniques have to be used. We propose to use the bootstrap method for that purpose. The main idea of bootstrap is to replace an unknown distribution by the respective empirical distribution function. Then, for a predefined number of bootstrap iterations B , a subsample from the empirical distribution function is drawn. In our case, for a single bootstrap iteration, n persons are drawn from the original sample with replacement. The proposed procedure is applied to the sampled data set, including the model selection using cross-validation. Therefore, the additional variance originating from the process of model selection is incorporated in the resulting confidence intervals. Finally, for every parameter bootstrap confidence intervals can be calculated using the empirical $\alpha/2$ and $1 - \alpha/2$ quantiles from the B bootstrap estimates for the respective parameter.

7.4. Application to Pre-Election Data from Germany

The proposed method is applied to data from the German Longitudinal Election Study (GLES), see Rattinger et al. (2014). The GLES is a long-term study of the German electoral process. It collects pre- and post-election data for the several federal elections.

7.4.1. Data

The data we are using here originate from the pre-election survey for the German federal election in 2013. In this specific part of the study, the participants ($n = 1155$ after eliminating all incomplete cases) were asked to rank the most important parties (CDU/CSU, SPD, Greens, Left Party, FDP, we eliminated the smaller parties AfD and the Pirate Party) for the upcoming federal election on a scale from -5 to $+5$. Plass et al. (2015) used the data in the context of modelling approaches for undecidedness. The ranks Z_r reflect the general opinions of the participants of party r where $+5$ represents a very positive and -5 represents a very negative opinion. The main goal of this application is to analyse which characteristics of the participants are connected to the opinions of the single parties. For that purpose, we generated paired comparisons out of the rankings. A similar approach for the analysis of rank data using paired comparisons was proposed by Francis et al. (2010). They also discuss the advantages of a paired comparison approach to model this form of data. For each participant, the differences between the ranks of all parties were calculated, ending up with ordered paired comparisons with values between -10 and 10 . The response was narrowed down to an ordered response with five categories. The data now represent paired comparisons between all parties measured on an ordered five-point scale:

$$\begin{aligned}
 Z_r - Z_s \in \{6, 10\} &\mapsto Y_{(r,s)} = 1 : \text{"I strongly prefer party r over party s"} \\
 Z_r - Z_s \in \{1, 5\} &\mapsto Y_{(r,s)} = 2 : \text{"I slightly prefer party r over party s"} \\
 Z_r - Z_s = 0 &\mapsto Y_{(r,s)} = 3 : \text{"I have equal opinions of parties r and s"} \\
 Z_r - Z_s \in \{-5, -1\} &\mapsto Y_{(r,s)} = 4 : \text{"I slightly prefer party s over party r"} \\
 Z_r - Z_s \in \{-10, -6\} &\mapsto Y_{(r,s)} = 5 : \text{"I strongly prefer party s over party r"}
 \end{aligned}$$

Within the GLES study, several characteristics of the participants are observed that possibly could affect the preference for the single parties. For our application, the following covariates are considered:

- Age: age of participant in years
- Gender: female (1); male (0)

- East Germany: East Germany/former GDR (1); West Germany/former FRG (0)
- Personal economic situation: good or very good (1); neither/nor, bad or very bad (0)
- School leaving certificate: Abitur/A levels (1); else (0)
- Unemployment: currently unemployed (1); else (0)
- Attendance in Church/Mosque/Synagogue/...: at least once a month (1); else (0)
- Have you been a German citizen since birth: yes (1); no (0)

7.4.2. Results

In the following, the results for the proposed method are presented for a model where all covariates described above are considered as possibly influential variables. The optimal model is determined by 10-fold cross-validation. Figure 7.2 shows the deviances obtained

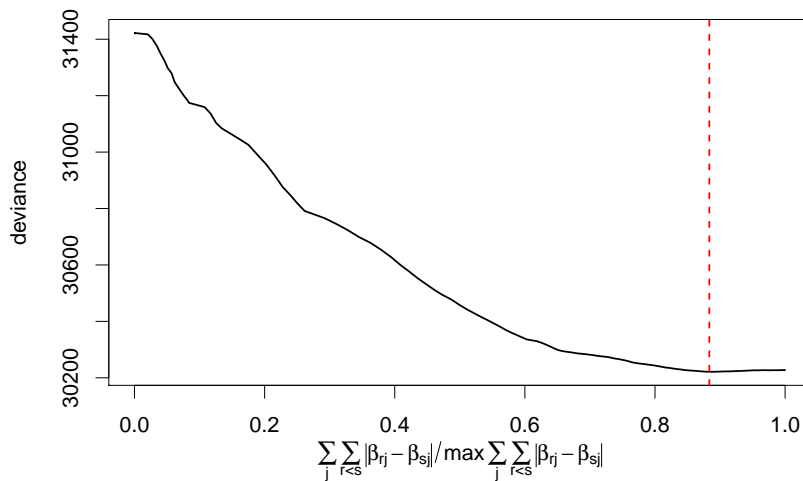


Figure 7.2.: Deviance path for 10-fold cross-validation, dashed vertical line represents model with lowest deviance.

by cross-validation plotted against the (normed) size of the penalized differences. Strong penalization corresponds to values close to 0, weak penalization to values close to 1. The dashed vertical line represents the model with the lowest deviance. Figure 7.3 shows the corresponding coefficient paths for the threshold parameters θ_1 and θ_2 and the party-specific intercepts $\beta_{10}, \dots, \beta_{m0}$. These parameters are not penalized. In principle, they might be

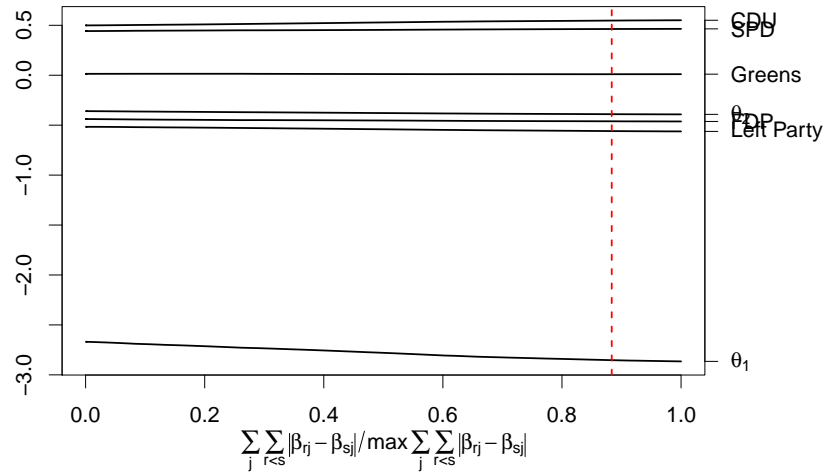


Figure 7.3.: Coefficient paths for all unpenalized parameters (threshold parameters θ_1 and θ_2 and party-specific intercepts). Dashed vertical line represents optimal model according to 10-fold cross-validation.

different for different tuning parameters λ . In the current application, it is seen that both the threshold parameters and the intercepts hardly change along their paths.

Figure 7.4 shows the corresponding coefficient paths for the eight covariates. The coefficient paths are drawn separately for each covariate. It is seen how the penalty term enforces clustering of the different parties. The dashed vertical lines represent the optimal model according to the 10-fold cross-validation.

The coefficient paths allow for interesting insights into how the preference of the voters for certain parties depends on characteristics of the voters themselves. Let us first consider the covariate unemployment. With respect to unemployment, the parties can be divided into two main clusters. The Left party and the Greens in one cluster, CDU, SPD and FDP in another cluster. As a global tendency one sees that unemployed persons tend to prefer the younger parties (Greens and Left Party) while the tendency to the more established parties (SPD, CDU, FDP) is reduced. In the optimal model, the second cluster of parties can be further divided into a cluster of SPD and FDP and a cluster only consisting of CDU. For gender, four different clusters are identified in the final model. The Greens are much more attractive for female than for male voters and form a cluster of their own. The SPD and the Left party seem almost equally attractive for males and females while the CDU and the FDP are more attractive for males. For the variable school leaving certificate, a very sparse solution with only two clusters (Greens vs. all other parties) emerged confirming the reputation of the Greens to be a party for academics. The German citizenship was

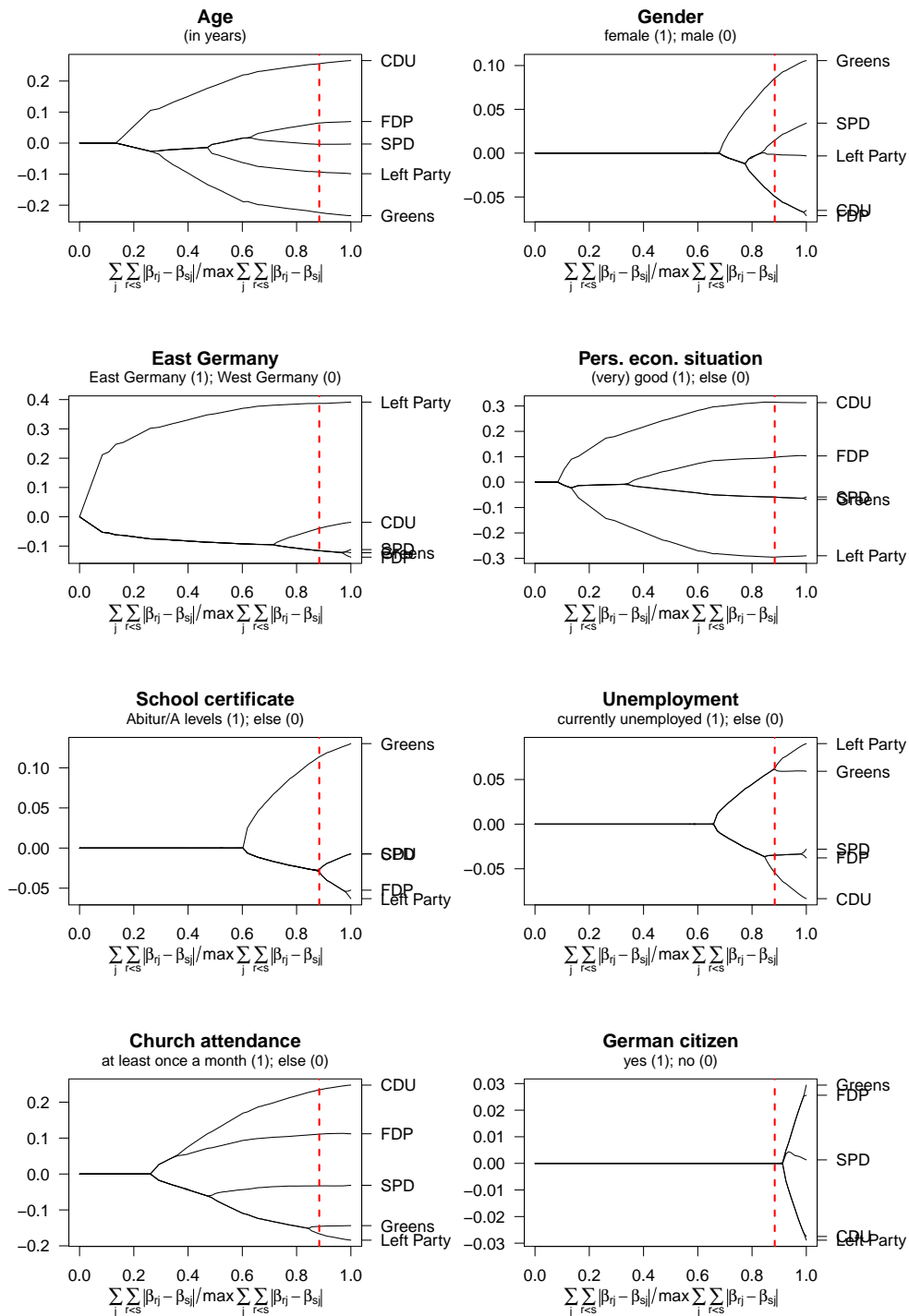


Figure 7.4.: Coefficient paths separately for all eight covariates. Dashed vertical lines represent optimal model according to 10-fold cross-validation.

completely eliminated from the model, naturalized citizens do not systematically prefer

other parties than citizens that were German citizens since birth. The variables age and church attendance have a specific impact on the preference of parties and every party forms a cluster of its own.

Similar to the results in chapters 4 and 5, again effect stars (see Appendix A for more details) can be used to visualize the results. As seen in Figure 7.4, the coefficients can be grouped by covariates. Therefore, per covariate one effect star can be plotted to visualize the effect of the respective covariate for all parties. Figure 7.5 shows the respective effect stars

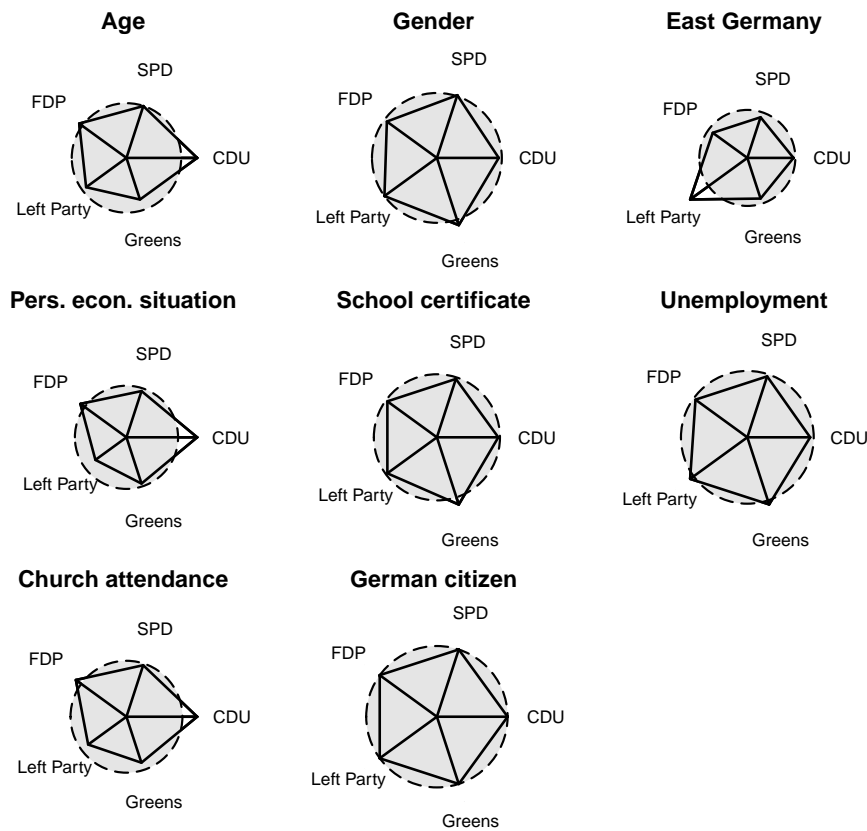


Figure 7.5.: Effect stars for all eight covariates illustrating parameter estimates corresponding to cv-optimal model.

illustrating the estimates corresponding to the optimal model found by cross-validation. As usual in effect stars, the lengths of the rays correspond to the exponentials of the respective estimates. The dashed circle has a radius equal to $\exp(0) = 1$ and, therefore, represents effects equal to zero. In most cases, the estimates vary around the circle. Some effects are outside the circle representing positive effects and some effects are within the circle representing negative effects. The covariate German citizenship is eliminated from the model and, therefore, all effects are located exactly on the no-effect circle. In contrast to Figure 7.4, effect stars only show the actual estimates instead of the complete paths. This

allows for a better comparability between the effects of different covariates. A disadvantage of effect stars is that the cluster effect of the penalty is not displayed anymore.

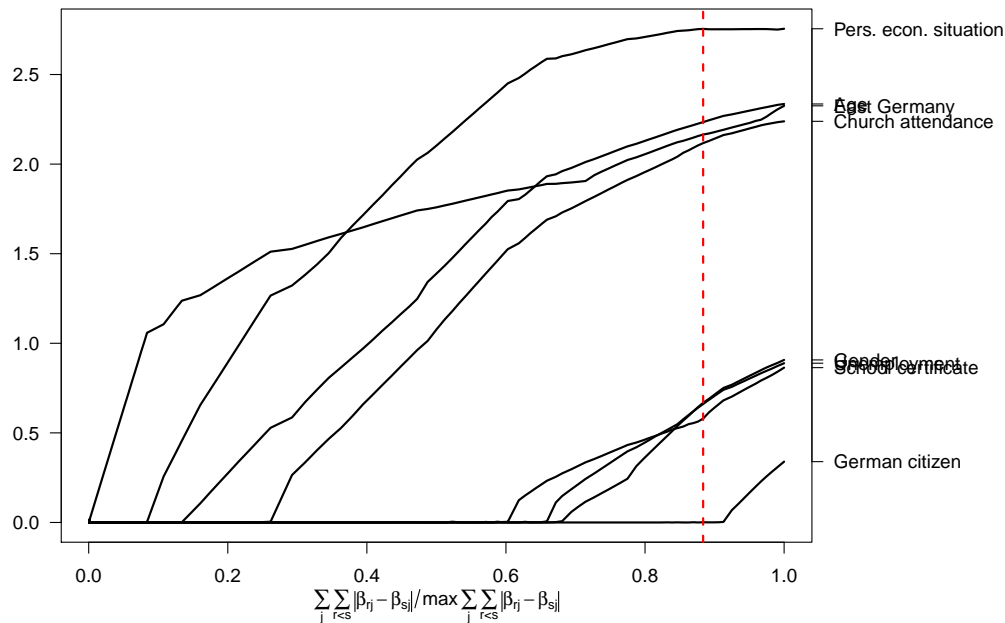


Figure 7.6.: Paths representing the sums of absolute differences for all eight covariates. Dashed vertical line represents optimal model according to 10-fold cross-validation.

Figure 7.6 shows the paths for whole covariates represented by the sum of absolute differences between all parameters corresponding to one covariate. Every covariate is represented by a single path. With the used penalty term, the sum of the absolute differences between all parameters corresponding to one covariate can be seen as a measure of effect strength for this covariate. Again, one has to keep in mind that all covariates have been standardized. It can be seen that, not very surprisingly, the personal economic situation of the voters is the most important modifier of the preference of a party in the data set. Yet, the first covariate that is included (for decreasing tuning parameter λ) is the covariate East Germany. Even 23 years after the German reunification, the differences between the former GDR and the former FRG were still extremely relevant in 2013. Also the covariates age and church attendance have very strong effects. Again, it can be seen that the variable German citizenship since birth is eliminated from the model. Figure 7.6 can provide valuable additional information on the paths depicted in Figure 7.4 where the variable importance is harder to recognize due to the different scales in the single plots.

Finally, $B = 500$ bootstrap iterations were performed to receive confidence intervals. Figure 7.7 depicts the estimates of all (penalized) parameters together with the corresponding

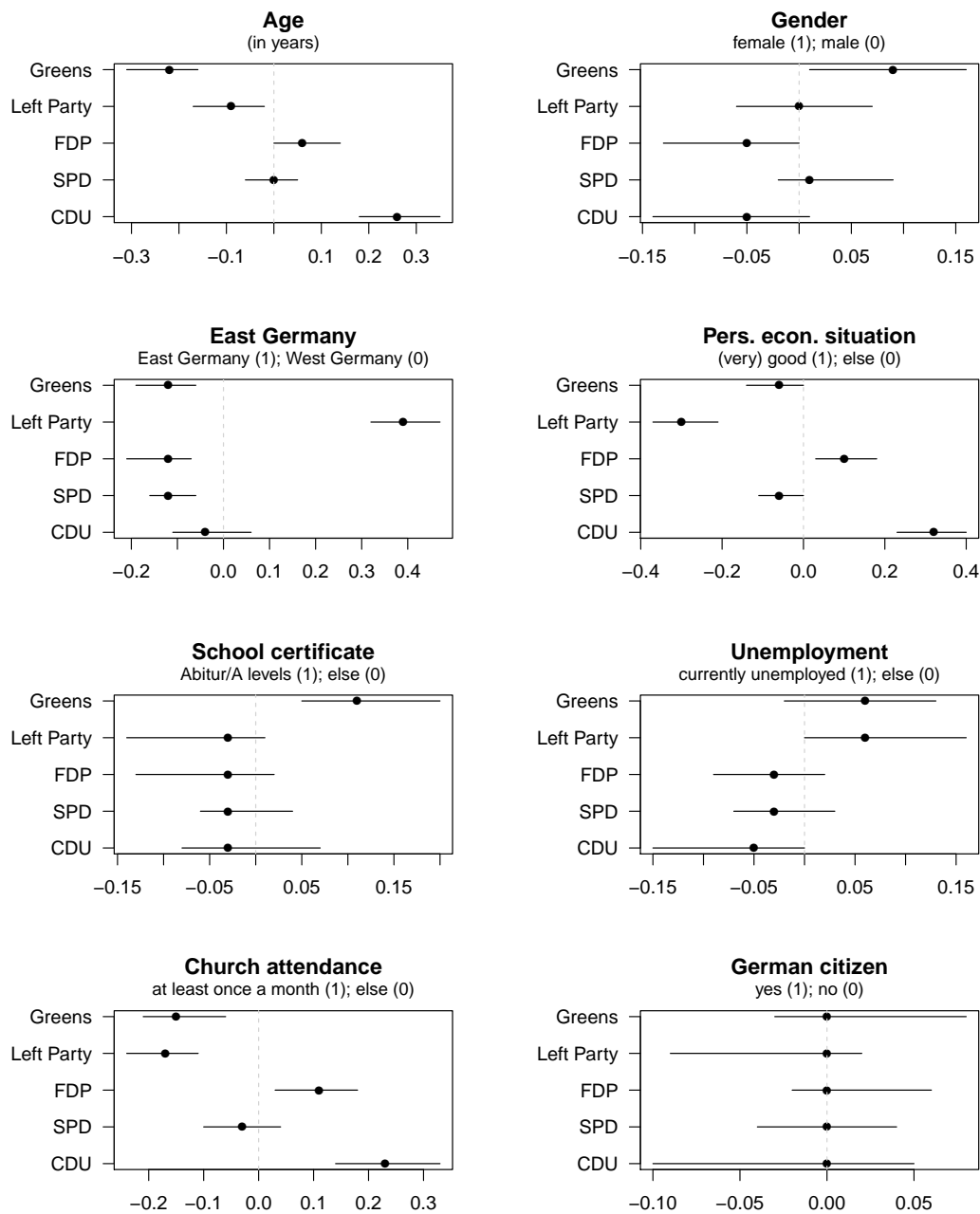


Figure 7.7.: Parameter estimates and 95% bootstrap confidence intervals separately for all eight covariates.

95% bootstrap confidence intervals. It can be seen if two clusters differ significantly from each other. For example, the parameters for the Left party and the Greens are not significantly different for church attendance although they are splitted into two different clusters. For the covariate unemployment, no parameter is significantly different from zero although three different clusters were estimated. Except for the covariates unemployment and Ger-

7.4.3. Inclusion of Twofold Interactions

Figure 1 is a line plot showing the distribution of the normalized difference in coefficients, $\sum_{i \leq s} |\beta_{ij} - \beta_{sj}| / \max_{i \leq s} \sum_{i \leq s} |\beta_{ij} - \beta_{sj}|$, for various combinations of variables and regions. The x-axis ranges from 0.0 to 1.0, and the y-axis ranges from 0.0 to 3.0. A vertical dashed red line is at x=0.5. The plot shows many curves, with some rising sharply from the origin and others rising more gradually. The curves represent different combinations of variables and regions, as indicated by the legend on the right.

Legend:

- Pers. econ. situation
- Church attendance
- Age
- East Germany
- East Germany : Church attend.
- Pers. econ. situation : Unemployment
- School certif. : Church attend.
- Age : East Germany
- Pers. econ. situation : German citizen
- School certif. : Unemployment
- Gender : East Germany
- Gender : School certif.
- Pers. econ. situation : German citizen
- Gender : East Germany
- East Germany : Pers. econ. situation

the absolute differences for all influence variables. Similar to the main effects model, the covariates personal economic situation, church attendance, age and East Germany are the most important influence variables. Yet, there are also some important interactions, like for example the interaction between East Germany and church attendance. In total, 13 out of all possible 28 influence variables are eliminated completely from the model. The labels in Figure 7.8 only show the names of the influence variables which entered the model.

Again, 95% bootstrap confidence intervals ($B = 500$) were calculated for all parameters (not shown here). Figure 7.9 shows all twofold interactions where at least one parameter was

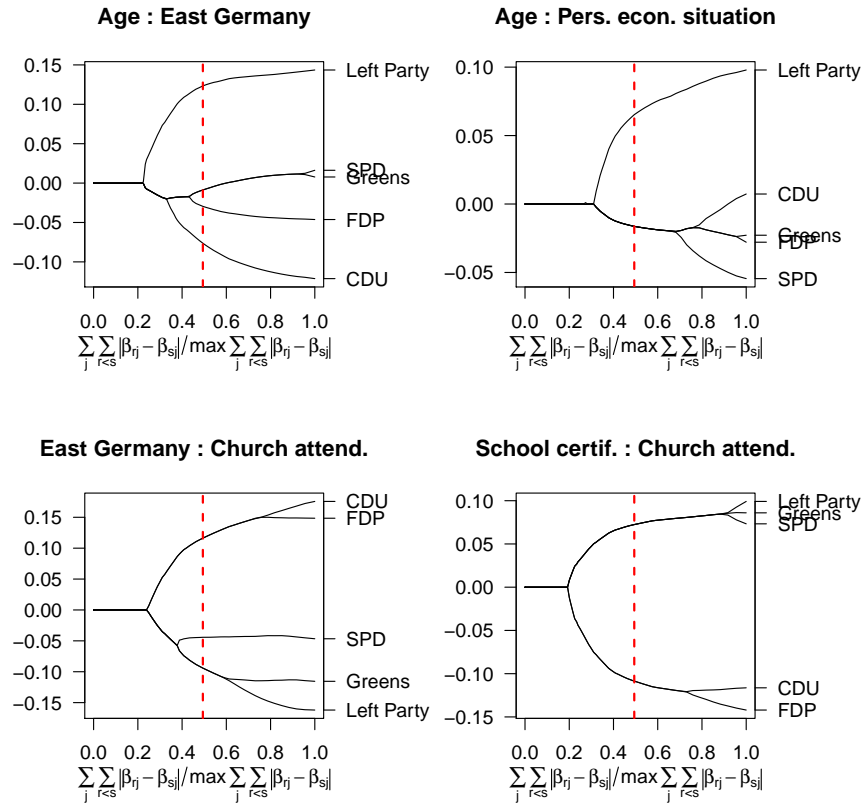


Figure 7.9.: Coefficient paths separately for all eight covariates and for all significant twofold interactions. Dashed vertical lines represent optimal model according to 10-fold cross-validation.

significantly different from zero. The paths of the main effects are very similar to Figure 7.4 and are left out for the sake of brevity. Except for the covariates German citizenship and unemployment, all main effects were significant (i.e. at least one parameter was significant different from zero). The interactions allow for additional insights into how covariates affect the preferences for the single parties. Exemplarily, we examine the interaction between age and the personal economic situation. Following the main effects, the preference for the Left party is decreased with growing age and if voters are in a good economic situation. After all, the interaction between both covariates has a negative effect. Therefore, for voters in a good economic situations the preference for the Left party increases with growing age.

7.5. Concluding Remarks

A model that explicitly accounts for heterogeneity in (possibly ordered) paired comparison models is proposed. The heterogeneity is modeled by the incorporation of subject-specific covariates. The model is estimated using a specific L_1 -type penalty. The penalty has two

main features: First, the penalty clusters items with regard to certain covariates. Therefore, one can identify clusters of items whose preferences are equally affected by a covariate. Second, the penalty can eliminate whole covariates from the model indicating that the respective covariates do not affect the preference for one or another item. Bootstrap intervals can be calculated which can be used to check if certain parameters differ significantly.

In particular the ability to select and cluster distinguishes the method from the few methods that are able to include covariates in paired comparison models. Francis et al. (2010) and Francis et al. (2002) include covariates but do not select the relevant ones, Casalicchio et al. (2015) presented a boosting approach that is able to select explanatory variables but is unable to detect clusters. Moreover, an advantage of penalty methods over boosting approaches is that the structure of the regularization is more clearly defined. In contrast to Strobl et al. (2011), where the underlying structure is searched for by recursive partitioning techniques, we consider a parametric model that allows for easy interpretation of parameters and clustering.

The proposed method could be extended in various ways. First, the restriction of the covariate effects to linear terms could be weakened by allowing for smooth covariate effects. A big challenge with such an approach would be to find an appropriate penalty term to have a similar cluster effect as for the linear terms. Second, the model could be extended by item-specific covariates similar to Chapter 8. For the application to the data from the GLES in this chapter, this would correspond to the inclusion of party-specific covariates, for example the popularity of the respective leading candidates.

8. Extended Ordered Paired Comparison Models with Application to Football Data

8.1. Introduction

Bayern München has been the dominating team in the season 2012/2013 of the German football league Deutsche Bundesliga. The dominance can be seen from the ranking according to the final points order. In the Bundesliga the winning team gains 3 points, the losing team receives nothing, and both teams gain 1 point if the match is drawn. This scheme of distributing points according to the outcome of the match can be seen as an ad hoc measure of the strengths of teams. But it is not without problems. In particular, if a team wins it is irrelevant if the adversary was a weak or a strong team. In the same way, each team gains one point in a draw, although, if the difference in strengths is large, the performance is weak for the stronger team but strong for the weaker team. A more elaborate way to measure the strength of teams is by considering the strength of a team as a latent trait and the performance, that is, the observable results, as determined by the latent traits of both teams. Models of this type have some tradition in statistics, in particular Bradley-Terry (BT-) models have been used to model competitions. Proposed by Bradley and Terry (1952), the model has been widely used to measure underlying strength in sport competitions. Dynamic models were considered, for example, by Fahrmeir and Tutz (1994), Knorr-Held (2000), Glickman and Stern (1998), and, more recently, by Cattelan et al. (2013).

In this chapter, we analyse the results of the German Bundesliga. We use a general latent trait model that does not only account for draws but allows for ordinal response categories that represent the competition results, thereby aiming at the efficient use of the information in the data. The model also includes an effect that represents the advantage in playing at home, which can also vary over teams. Aspects of the model have been already proposed in the literature. Models that allow for a draw were proposed by Rao and Kupper (1967),

This chapter is a modified version of Tutz and Schauburger (2015a), previous work on the issue can be found in the technical report 151 (Tutz and Schauburger, 2014). See Chapter 1 for more information on the personal contributions of all authors and textual matches.

Davidson (1970) and used to model sports tournaments by Cattelan et al. (2013). Models that allow for any number of ordered response categories were proposed by Tutz (1986) and Agresti (1992). Heterogeneity of the home advantage has been considered by Kuk (1995), Knorr-Held (2000), and Glickman and Stern (1998), but only for models with a draw. An approach to find clusters of teams that can not be distinguished has been proposed by Masarotto and Varin (2012). Here, it is extended to work in the general model and also to find clusters of teams with the same home advantage.

In a second step it is investigated how much of the variation in the strengths of the teams is explained by team-specific covariates. It is especially interesting how much of the strength of a team is explained by the budget. Is Bayern Munich the best team because it is the richest club in Germany? For the analysis the estimated strength parameters are used and a model that includes effects of covariates is proposed. Estimation is based on penalization methods that allow to group the abilities of teams. We analyse the German Bundesliga data and demonstrate that the model with explanatory variables yields useful estimates.

In Section 8.2 we briefly describe the data. In Section 8.3 we introduce the general ordinal model and give results for the Bundesliga. Section 8.5 is devoted to the inclusion of team-specific explanatory variables.

8.2. German Bundesliga

Before defining latent trait models, which will be quite general for the modelling of competition results, we briefly describe the structure of the German Bundesliga competition. The tournament comprises $m = 18$ teams, we analyse the matches played in the 50th season of the Bundesliga from August 24, 2012 to May 18, 2013. The tournament structure is that of a double round-robin, each team competes twice against all the other teams, once on home ground and once away. On average, 42.5% of the matches were won by the home team, 25.5% of the matches ended with a draw and 32% of the matches were won by the away team. Table 8.1 shows the results ranked according to the final points order.

Two aspects from the final ranking are unique occurrences in the history of the Bundesliga. Bayern München was the dominating team for the season and set several new records. For example, Bayern München gained the highest number of points and victories for a team in one season. For the Spielvereinigung Greuther Fürth, it was the first participation in the German Bundesliga, they were the first team without a victory on home ground for a whole season.

	Points	Home	Away	Ability	QSE	Rank
FC Bayern München	91	44	47	2.562	0.377	1
Borussia Dortmund	66	33	33	1.361	0.314	2
Bayer 04 Leverkusen	65	39	26	0.983	0.306	3
FC Schalke 04	55	33	22	0.460	0.300	4
Eintracht Frankfurt	51	31	20	0.350	0.300	6
Sport-Club Freiburg	51	28	23	0.409	0.300	5
Hamburger SV	48	26	22	0.023	0.300	11
Borussia Mönchengladbach	47	29	18	0.235	0.300	7
Hannover 96	45	32	13	0.074	0.300	9
1. FC Nürnberg	44	27	17	0.057	0.300	10
VfB Stuttgart	43	19	24	-0.183	0.302	13
VfL Wolfsburg	43	17	26	0.000	0.300	12
1. FSV Mainz 05	42	26	16	0.084	0.300	8
SV Werder Bremen	34	20	14	-0.272	0.303	14
FC Augsburg	33	20	13	-0.562	0.307	16
1899 Hoffenheim	31	19	12	-0.616	0.308	17
Fortuna Düsseldorf	30	21	9	-0.287	0.303	15
SpVgg Greuther Fürth	21	4	17	-0.956	0.315	18

Table 8.1.: Final ranking of the German Bundesliga 2012/2013 including points in home matches and away matches; the last three columns show the estimated abilities, quasi standard errors and the ranking corresponding to the estimated abilities for the ordered model including a home advantage parameter

8.3. Ordered Paired Comparison Model with Home Advantage

In the following, latent trait models are considered. The basic concept is that the probability of winning or losing is determined by the underlying strengths of teams. While the strengths are fixed the result of a competition is a random variable. The models can be used in all competitions where two teams or players compete in a tournament like tennis, football, and chess. In some sports there is a clear winner, in others draws can occur. Another feature that depends on the form of competition is that home effects can occur. In particular, in football playing at the home ground seems to be advantageous. We will consider a general model that can account for all these effects.

8.3.1. The Basic Binary Bradley-Terry Model

Let $\{a_1, \dots, a_m\}$ denote the set of teams or players that compete. In the simplest case when a team can only win or lose the relation between the underlying strengths of the teams and the outcome can be modeled by the Bradley-Terry model (Bradley and Terry, 1952), which specifies for the probability that a_r beats a_s

$$P(a_r \succ a_s) = \frac{\exp(\gamma_r - \gamma_s)}{1 + \exp(\gamma_r - \gamma_s)}.$$

The parameters $\gamma_r, r = 1, \dots, m$, can be interpreted as the strengths of the teams $\{a_1, \dots, a_m\}$. For $\gamma_r = \gamma_s$ the probability that a_r wins against a_s is 0.5, for growing distance $\gamma_r - \gamma_s$ the probability increases accordingly.

With the random variable $Y_{(rs)} = 1$ if $r \succ s$ and $Y_{(rs)} = 0$ otherwise one obtains the logit model

$$\log \left(\frac{P(Y_{(rs)} = 1)}{P(Y_{(rs)} = 0)} \right) = \gamma_r - \gamma_s.$$

The model in this form is not identifiable because strengths parameters $\gamma_r + c$ for fixed value c yield the same probabilities. Therefore, a constraint is needed. We choose to fix one parameter, that is, γ_m is set to zero defining object a_m to be the reference object. In our case the reference team is Wolfsburg.

8.3.2. Ordinal Models Including the Advantage in Playing at Home

Let now the success of team a_r in a match between team a_r and a_s be measured on an ordinal scale represented by $Y_{(rs)} \in \{1, \dots, K\}$, for odd K , where low numbers denote dominance of team a_r and high numbers dominance of team a_s . The scale is assumed to be symmetric regarding the two teams. That means the numbers 1 to K represent categories like "strong dominance of team a_r ", "weak dominance of team a_r ", "draw", "weak dominance of team a_s ", "strong dominance of team a_s ". In the simplest case, where $K = 3$, the responses are "team a_r wins", "draw", "team a_s wins". But to exploit the information contained in the results of matches one might also consider the differences in scored goals as indicators of dominance. In the application we use a difference of at least 2 goals as an indicator for strong dominance and work with a 5-point scale. A model that allows for ordered responses is the cumulative type model

$$P(Y_{(rs)} \leq k) = F(\eta_{(rs)k}), \quad \eta_{(rs)k} = \theta_k + \gamma_r - \gamma_s, \quad (8.1)$$

where $F(\cdot)$ is a symmetric distribution function, which in Bradley-Terry type models is the logistic distribution function. The linear predictor $\eta_{(rs)k}$ contains the difference in strengths $\gamma_r - \gamma_s$ and so-called threshold parameters that account for the frequency of the response categories. The symmetry of the response categories entails the restrictions $\theta_k = -\theta_{K-k}$, $t = 1, \dots, [K/2]$. That means, in particular, that for teams with identical strengths, $\gamma_s = \gamma_r$, one obtains $P(Y_{(rs)} = k) = P(Y_{(rs)} = K + 1 - k)$. For the most important case $K = 3$ one obtains $P(Y_{(rs)} = 1) = P(Y_{(rs)} = 3)$, that means that the probability of winning is the same for both teams. Similar restrictions are needed if the number of response categories

K is even, which is relevant only in competitions that do not allow for a draw (see Tutz (1986)).

The cumulative model (8.1) is able to use the information contained in ordered responses; with more categories better estimates are to be expected. In the literature alternative models have been proposed. In particular, the adjacent category model, proposed by Agresti (1992) is an alternative that also uses the full information in ordinal data. It is an extension of the three category model of Davidson (1970), which can also be estimated within a log linear model framework (Dittrich et al., 2004). Further applications of the adjacent categories model are found in Dittrich et al. (2000), Böckenholt and Dillon (1997a) and Böckenholt and Dillon (1997b).

Home Effects

When modelling competitions one also has to account for the advantage deriving from playing at home. Let the first index of response $Y_{(rs)}$ represent the home team. To include the advantage of the home team, the linear predictor is extended to

$$\eta_{(rs)k} = \delta + \theta_k + \gamma_r - \gamma_s,$$

where δ represents the home effect. It is typically positive and, therefore, increases the probability for low response categories that correspond to the dominance of team a_r . It is easily derived that for $K = 3$ and equal strength, $\gamma_r = \gamma_s$, δ reflects the proportion of odds for winning of team a_r and winning of team a_s ,

$$\delta = \frac{1}{2} \log \left(\frac{P(Y_{(rs)} = 1)/(1 - P(Y_{(rs)} = 1))}{P(Y_{(rs)} = 3)/(1 - P(Y_{(rs)} = 3))} \right).$$

However, it is questionable that the home effect is the same for each team. Some teams may profit more from playing at home than others. A team-specific home effect is obtained by using the predictor

$$\eta_{(rs)k} = \delta_r + \theta_k + \gamma_r - \gamma_s.$$

In this general model the γ -parameters do not represent the strengths of teams per se because performance depends on whether playing at home or not. Again, for $K = 3$ and equal strength, $\gamma_r = \gamma_s$, the home effect when playing at the home ground of team a_r is given by the proportion of odds for winning (of team a_r) against losing

$$\delta_r = \frac{1}{2} \log \left(\frac{P(Y_{(rs)} = 1)/(1 - P(Y_{(rs)} = 1))}{P(Y_{(rs)} = 3)/(1 - P(Y_{(rs)} = 3))} \right).$$

But in the general model, the proportion of odds for winning (of team a_r) against losing when playing at the home ground of the second team a_s are not just the inverse of the proportion when playing at the home of team a_r as in the model with constant home effect.

By defining $\tilde{\gamma}_r = \delta_r + \gamma_r$, the predictor obtains the form $\eta_{(rs)k} = \theta_k + \tilde{\gamma}_r - \gamma_s$. As in the basic model (8.1), the result of a match is determined by the difference of strength, but now it is $\tilde{\gamma}_r - \gamma_s$. Therefore, $\tilde{\gamma}_r$ represents the strength when playing at home and γ_r the strength when not playing at home.

8.3.3. Fitting the Model

Estimation of the cumulative model can be embedded into the framework of generalized linear models (GLMs), which were thoroughly investigated by McCullagh and Nelder (1989). For data $Y_{(rs)} \in \{1, \dots, K\}$, $r, s \in \{1, \dots, m\}$ the linear predictor can be written as

$$\eta_{(rs)k} = \delta_r + \theta_k + \gamma_r - \gamma_s = \delta_r + \theta_k + x_2^{(r,s)}\gamma_2 + \dots + x_m^{(r,s)}\gamma_m = \delta_r + \theta_k + (\mathbf{x}^{(r,s)})^T \boldsymbol{\gamma},$$

where the components of the $(m-1)$ -vector $\mathbf{x}^{(r,s)}$ are given by

$$x_j^{(r,s)} = \begin{cases} 1 & j = r \\ -1 & j = s \\ 0 & \text{otherwise.} \end{cases}$$

Thus, it is a cumulative model with threshold θ_k , the additional parameter δ_r and "predictor" $\mathbf{x}^{(r,s)}$. The predictor can also be given by $\mathbf{x}^{(r,s)} = \mathbf{1}_r - \mathbf{1}_s$, where $\mathbf{1}_r = (0, \dots, 0, 1, 0, \dots, 0)$ has length $m-1$ with 1 at position r . Cumulative models have been considered in particular by McCullagh (1980), estimation within the framework of multivariate GLMs was considered by Fahrmeir and Tutz (2001) and by Tutz (2012). The embedding into this framework allows to use the familiar goodness-of-fit statistics as well as likelihood ratio statistics to test hypotheses if one assumes that the observations given the abilities are independent.

8.3.4. Football Data

We first consider the modelling of the football data under the assumption that the home advantage is global, that is, it does not depend on the team. Then, one has one strength parameter for each team and does not have to distinguish between the strength when playing at home or away. In the following, we try to use the available information by using a 5-point scale to evaluate the performance in a competition. The categories refer to "winning with

a difference of at least two goals", "winning with a difference of less than two goals" and "draw" as the middle category.

Global Home Effect Model

The estimated home advantage is $\hat{\delta} = 0.293$; for the threshold parameters one obtains $\hat{\theta}_1 = -\hat{\theta}_4 = -1.66$ and $\hat{\theta}_2 = -\hat{\theta}_3 = -0.65$. If one assumes that two teams have equal abilities, the threshold parameters correspond to probabilities of 0.41 for a victory of the home team, 0.31 for a draw and 0.28 for a victory of the away team. Thus the home advantage can definitely not be ignored. The tendency is also seen from the averages over all games, because 42.5% of the matches were won by the home team, 25.5% of the matches ended with a draw and 32% of the matches were won by the away team. But these numbers are averages over games played by teams with differing abilities. The strength of the latent trait model is that the home advantage takes this variation of abilities into account when estimating the home advantage. Table 8.1 shows the estimated abilities together with the ranks according to the final points. It is seen that for the best teams the rank is in accordance with the estimated abilities but in the middle part of the table there are some permutations. However, quasi standard errors, computed following Firth and De Menezes (2004) suggest that the permutations are not to be taken too seriously. This will be investigated in more detail in Section 8.4.

Team-Specific Home Effects

The question if home effects are team-specific is investigated by computing the likelihood ratio test for the hypothesis that all effects are equal, yielding a value of 24.69 on 17 degrees of freedom, which corresponds to a p -value of 0.102. Therefore it is not significant when using significance level 0.05, but nevertheless it is small. If one uses a 3-point scale that only distinguishes between "winning", "draw" and "losing", the p -value is 0.022, which is definitely smaller. In Table 8.2 the estimates and the corresponding ranks are given when one distinguishes between home and away strength. As always in the applications we use the more informative 5-point scale. It is seen that for the best performers the order is very stable. It is the same when playing at home or away or when not distinguishing between the two. But one also finds large differences. For example, Hannover has rank 4 at home, but rank 17 when playing away with a difference of 1.167 in abilities. For Wolfsburg the ranks are just the opposite, it has rank 17 at home and rank 4 when playing away. Wolfsburg is also one of the few teams that have larger ability when playing away with a negative value for the home effect.

	Overall		Home		Away	
	Ability	Rank	Ability	Rank	Ability	Rank
FC Bayern München	2.562	1	1.871	1	2.220	1
Borussia Dortmund	1.361	2	0.901	2	0.729	2
Bayer 04 Leverkusen	0.983	3	0.851	3	0.013	3
FC Schalke 04	0.460	4	0.191	6	-0.505	6
Eintracht Frankfurt	0.350	6	0.258	5	-0.782	11
Sport-Club Freiburg	0.409	5	-0.068	8	-0.334	5
Hamburger SV	0.023	11	-0.490	11	-0.708	8
Borussia Mönchengladbach	0.235	7	-0.020	7	-0.722	9
Hannover 96	0.074	9	0.338	4	-1.505	17
1. FC Nürnberg	0.057	10	-0.140	9	-0.999	14
VfB Stuttgart	-0.183	13	-1.024	16	-0.564	7
VfL Wolfsburg	0.000	12	-1.262	17	0.000	4
1. FSV Mainz 05	0.084	8	-0.293	10	-0.782	10
SV Werder Bremen	-0.272	14	-1.014	15	-0.803	12
FC Augsburg	-0.562	16	-0.881	13	-1.541	18
1899 Hoffenheim	-0.616	17	-0.966	14	-1.486	16
Fortuna Düsseldorf	-0.287	15	-0.695	12	-1.204	15
SpVgg Greuther Fürth	-0.956	18	-2.278	18	-0.906	13

Table 8.2.: Comparison of the estimated abilities from the model with a global home advantage to the estimated abilities from the model with team-specific home advantages

Ranks and Abilities

The traditional measure for the performance of teams is the number of gained points summarized over all games. It is interesting to investigate, how this measure that is defined by the association of the football league is related to the abilities found by the fitting of a latent trait model. It turns out that the correlation is quite high. For the 50th season we obtained a correlation of 0.982, which means that gained points and abilities measure almost the same. One may wonder if this is an effect of the specific scheme, which gives winning team 3 points, the losing team nothing, and both teams 1 point if the match is drawn. Is this scheme appropriate under the assumption that the latent trait model is an adequate representation of the link between the observations and the latent abilities? Therefore, we shortly investigate how the scheme of distributing points influences the correlation between number of points and estimated abilities. In a general scheme, the winning team gains $w > 0$ points, the losing team nothing and both teams $d > 0$ points if the match is drawn. It is easily derived that for constant proportion w/d one obtains up to a scaling factor the same number of points. Because a scaling factor is irrelevant when computing the correlation, it suffices to vary only one of the two parameters w and d . Without loss of generality we set $d = 1$. Figure 8.1 shows the dependence of the correlation on the gained points for winning w (bold faced curve). It is seen that the maximum is obtained for $w = 2.2$, which is not far from the 3 points fixed in the regulations. The curve decreases slowly beyond its maximum. That means, also much higher points could be given to the

winning team and still the number of points would be in strong accordance with the abilities. The strong correlation found for the data could be related to the double round robin structure of the tournament. In paired comparisons, where not all pairs are evaluated, we expect lower correlations. To investigate the effects we have drawn sub samples of the paired comparisons containing 50% of the pairs. Two specific sub samples are the results of the first round and the second round. Figure 8.1 shows the corresponding correlations. It is seen that, depending on the sample, correlations can be much smaller. That means, in particular, for an ongoing season, when not all matches have been played, the ranking by points and abilities are less strongly connected.

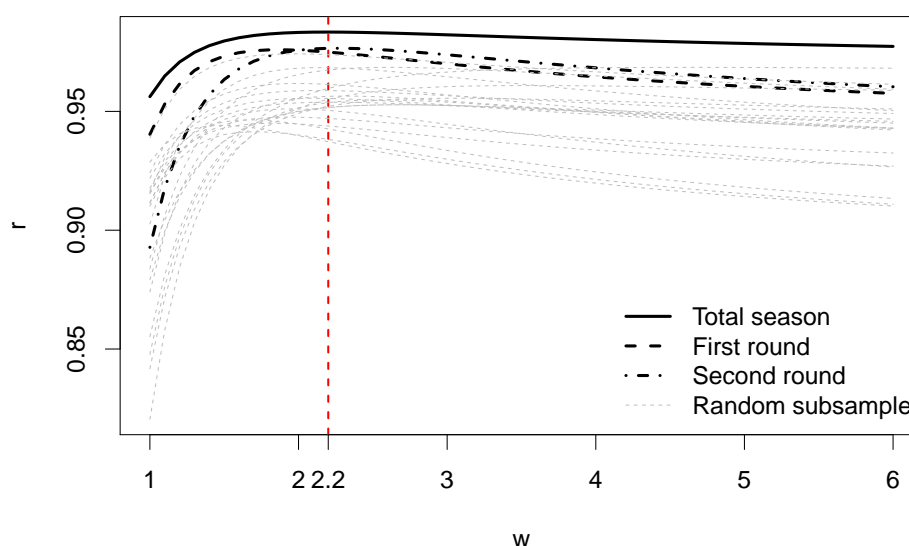


Figure 8.1.: Correlations plotted against the points gained when winning for the whole season (bold faced curve), first round (dashed), second round (dashed dotted) and several sub samples.

8.4. Identification of Clusters

A disadvantage of simply measuring the performance of teams by points is that there is no information on the precision of this measurement tool. In contrast the latent trait model allows to evaluate which teams are really to be distinguished. One way is to consider the standard errors, which contain the information about the relevance of differences between the estimated abilities. An alternative approach is to explicitly aim at finding clusters of teams which share the same ability by using regularization techniques. Clustering techniques proposed by Bondell and Reich (2009) and Gertheiss and Tutz (2010) have been used

by Masarotto and Varin (2012) to cluster abilities in a paired comparison model which allows for draws. In this section we will use these techniques in the general case of ordinal response data. In Section 8.4.2 the method is extended to find clusters of abilities as well as clusters of home advantages.

8.4.1. Clustering of Teams

One way of obtaining regularized estimates is to use penalty terms that yield structured estimates. Instead of maximizing the log-likelihood, one maximizes the penalized log-likelihood

$$l_p(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \lambda J(\boldsymbol{\beta}),$$

where $l(\boldsymbol{\beta})$ denotes the familiar un-penalized log-likelihood, λ is a tuning parameter, and $J(\boldsymbol{\beta})$ is a penalty term. A specific penalty term, which enforces the clustering of abilities and which will also be useful later, is given by

$$J(\boldsymbol{\beta}) = \sum_{r < s} w_{rs} |\gamma_r - \gamma_s|, \quad (8.2)$$

where w_{rs} are specific weights. The penalty is a fusion type penalty, which enforces the fusion of abilities. By using the L_1 -norm it enforces, in particular, that for growing λ abilities are set equal. The effect of the penalty is also seen by looking at extreme values of the tuning parameter λ . If $\lambda \rightarrow \infty$, all strength parameters γ_r are estimated as identical.

Cluster	Ability
1 FC Bayern München	2.26
2 Borussia Dortmund	1.06
3 Bayer 04 Leverkusen	0.73
4 FC Schalke 04; Sport-Club Freiburg; Eintracht Frankfurt	0.01
5 Borussia Mönchengladbach; 1. FSV Mainz; Hannover 96; 1. FC Nürnberg; Hamburger SV; VfL Wolfsburg	0.00
6 VfB Stuttgart; SV Werder Bremen; Fortuna Düsseldorf	-0.04
7 FC Augsburg; 1899 Hoffenheim	-0.33
8 SpVgg Greuther Fürth	-0.70

Table 8.3.: Clusters of teams with corresponding abilities.

In the case of a global home advantage the procedure typically yields distinct clusters. Figure 8.2 shows the coefficient paths with the weights given by $w_{rs} = |\hat{\gamma}_r^{(\text{ML})} - \hat{\gamma}_s^{(\text{ML})}|^{-1}$, where $\hat{\gamma}_r^{(\text{ML})}$ denotes the maximum likelihood estimate of team a_r . For details of this weighting scheme, which yields more stable coefficient paths than un-weighted fusion penalties, see Gertheiss and Tutz (2010) and Masarotto and Varin (2012). The straight lines in Figure 8.2 represent the BIC (Schwarz, 1978) and the AIC (Akaike, 1973) criterion. Based on the BIC criterion one finds that the 18 teams are divided into eight clusters with abilities being

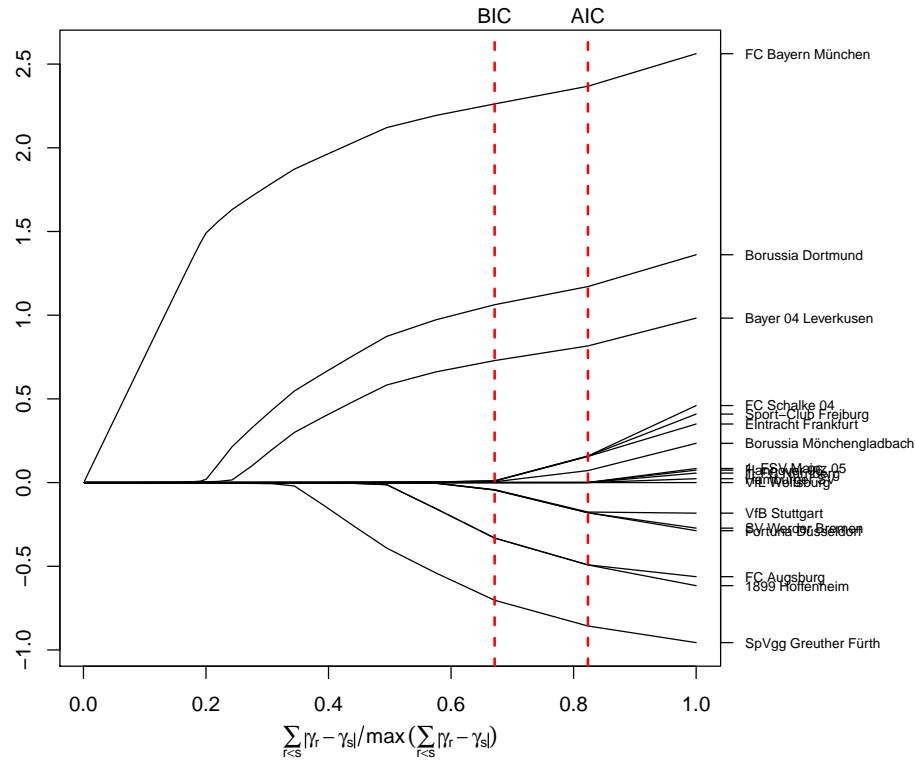


Figure 8.2.: Coefficient paths for ability parameters in the model with a global home advantage using an adaptive L_1 -penalty.

identical within clusters. Table 8.3 shows the clusters and the corresponding estimated abilities. It is seen that the three best teams and the worst team form clusters of their own. The second worst cluster contains two teams. All other teams are collected in three big clusters, which have rather similar abilities. In fact, if one measures abilities only up to one digit, they form just one big cluster.

8.4.2. Clustering of Teams and Home Effects

Clustering becomes much more difficult if one suspects team-specific home advantages because then one has to distinguish the strength when playing at home and the strength when playing away. A penalty term that clusters the home advantage, δ_r , the abilities when playing at home, $\gamma_r + \delta_r$, as well as the abilities when playing away, γ_r , is

$$J(\beta) = \sum_{r<s} w_{rs} |\gamma_r - \gamma_s| + \sum_{r<s} u_{rs} |\gamma_r - \gamma_s + \delta_r - \delta_s| + \sum_{r<s} v_{rs} |\delta_r - \delta_s|.$$

with $w_{rs} = |\gamma_r^{(ML)} - \gamma_s^{(ML)}|^{-1}$, $u_{rs} = |\gamma_r^{(ML)} - \gamma_s^{(ML)} + \delta_r^{(ML)} - \delta_s^{(ML)}|^{-1}$, and $v_{rs} = |\delta_r^{(ML)} - \delta_s^{(ML)}|^{-1}$. It enforces clustering of both abilities and the home advantage. For the selection of the optimal tuning parameter λ , we again use the BIC criterion

$$BIC(\lambda) = -2 \cdot l(\beta) + df(\lambda) \cdot \log(n),$$

where n is the number of observations. It depends on the degrees of freedom $df(\lambda)$ of the respective model. For penalized models, the degrees of freedom do not equal the number of parameters in the model because of the effects of shrinkage and variable selection. Therefore, following Buja et al. (1989), the degrees of freedom are calculated by $\text{tr}(2\mathbf{H} - \mathbf{H}^T\mathbf{H})$. Here, \mathbf{H} represents the hat matrix obtained in the last Fisher scoring step in the penalized iteratively re-weighted least squares (PIRLS) algorithm that is used. The algorithm and the corresponding hat matrix are described in more detail by Oelker and Tutz (2015).

Cluster (ability home)		Ability
1	FC Bayern München	1.84
2	Borussia Dortmund	0.61
3	Bayer 04 Leverkusen	0.44
4	Hannover 96; FC Schalke 04; Eintracht Frankfurt; Sport-Club Freiburg	-0.21
5	Borussia Mönchengladbach; 1. FC Nürnberg; 1. FSV Mainz 05; Hamburger SV	-0.22
6	Fortuna Düsseldorf	-0.48
7	FC Augsburg; SV Werder Bremen; VfB Stuttgart; 1899 Hoffenheim	-0.50
8	VfL Wolfsburg	-0.55
9	SpVgg Greuther Fürth	-1.56
Cluster (ability away)		Ability
1	FC Bayern München	1.68
2	Borussia Dortmund	0.17
3	Bayer 04 Leverkusen; VfL Wolfsburg	0.00
4	Sport-Club Freiburg; FC Schalke 04	-0.65
5	Borussia Mönchengladbach; VfB Stuttgart; Hamburger SV; Eintracht Frankfurt; 1. FSV Mainz 05; SV Werder Bremen; 1. FC Nürnberg; SpVgg Greuther Fürth	-0.66
6	Fortuna Düsseldorf	-0.91
7	Hannover 96; 1899 Hoffenheim; FC Augsburg	-0.94
Cluster (home advantage)		Ability
1	Hannover 96	0.73
2	Eintracht Frankfurt; Bayer 04 Leverkusen; 1. FC Nürnberg; Borussia Mönchengladbach; FC Schalke 04; FC Augsburg; 1899 Hoffenheim; 1. FSV Mainz 05; Fortuna Düsseldorf	0.44
3	Sport-Club Freiburg; Hamburger SV; Borussia Dortmund	0.43
4	SV Werder Bremen; FC Bayern München; VfB Stuttgart	0.15
5	VfL Wolfsburg	-0.55
6	SpVgg Greuther Fürth	-0.90

Table 8.4.: Clusters of teams when distinguishing between abilities when playing at home and playing not at home, and clusters of home advantages.

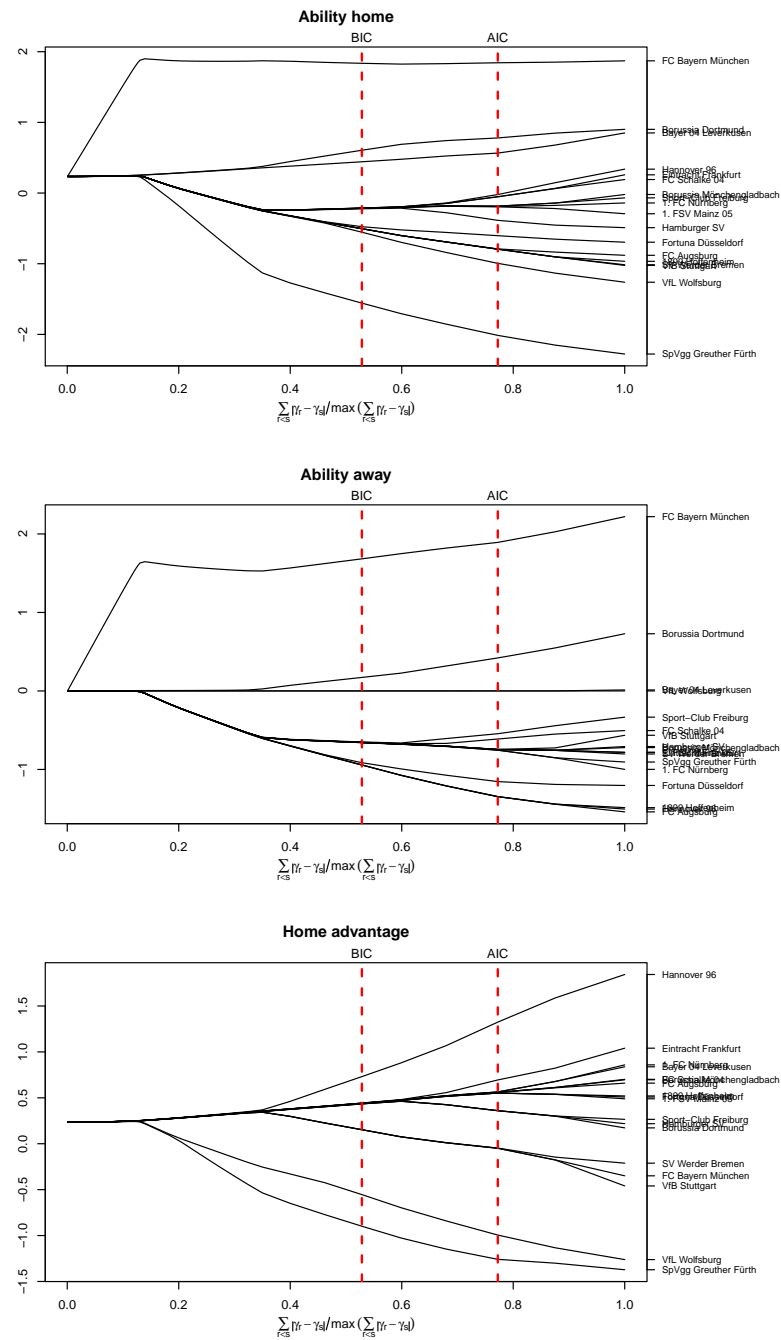


Figure 8.3.: Coefficient paths for home abilities, away abilities and home advantages in the model with team specific home advantages using an adaptive L_1 -penalty

Figure 8.3 shows the coefficient build-ups and Table 8.4 the corresponding clusters. For the strong teams one obtains very similar classes, but in particular in the middle different clusters are found when playing at home and away. Clustering of the home effect yields

essentially 5 classes; Hannover is a class of its own, the big clusters 2 and 3 are hardly different and there are even two clusters with negative home advantage.

8.5. Accounting for Explanatory Variables

Scaling of teams by use of paired comparison models yields estimated abilities but does not explain why some teams are better than others. If one wants to explain the variation in abilities, a natural way is to include covariates in the model. The most interesting variables are variables that characterize the clubs and, therefore, the teams, in contrast to variables that are shared by both teams like day of the week or weather when playing. Explanatory variables of the latter type are more interesting when items are compared and preference is to be modeled as a function of characteristics of the person that chooses. Explanatory variables of this type have been considered in Chapter 7 and, for example, by Dittrich et al. (1998) when modeling the preference for European universities.

8.5.1. A Model with Team-Specific Explanatory Variables

Let the data be given by $(Y_{(rs)}, r, s \in \{1, \dots, m\}, \mathbf{x}_1, \dots, \mathbf{x}_m)$ where $Y_{(rs)} \in \{1, \dots, K\}$ denotes the ordinal response and \mathbf{x}_r is a vector of explanatory variables linked to team a_r . Exemplarily, we will consider the budget of a club, which should be influential because the budget determines if a club is able to get the best and most expensive players.

In a general model that accounts for team-specific variables, the strength of the teams, γ_r , is replaced by $\gamma_r + \mathbf{x}_r^T \boldsymbol{\alpha}$ yielding the linear predictor

$$\eta_{(rs)k} = \delta_r + \theta_k + \gamma_r - \gamma_s + (\mathbf{x}_r - \mathbf{x}_s)^T \boldsymbol{\alpha}.$$

In this model, parameters are not identifiable because the parameters γ_r can not be distinguished from the parameters $\tilde{\gamma}_r = \gamma_r + \mathbf{x}_r^T \boldsymbol{\alpha}$. Therefore, additional constraints are needed to obtain unique estimates. A very restrictive model that is identifiable has been proposed by Springall (1973). He obtains identifiability by setting $\gamma_r = 0$, $r = 1, \dots, m$. The corresponding model assumes that the explanatory variables totally determine the abilities. It is hardly appropriate when a limited number of explanatory variables is available.

A much better and more flexible way to constrain estimates is to use a random effects model. By assuming that the strengths are random effects, for example, by assuming $\gamma_r \sim N(0, \sigma^2)$, parameters can be estimated within a random effects model, see Turner and Firth (2012) who used random effects models to account for correlations between responses. An alternative approach that is advocated here is to use penalized estimation procedures

within a fixed effects model framework. If one assumes that teams are clustered one can use the penalty (8.2). It penalizes the abilities that are not explained by covariates, γ_r , $r = 1, \dots, m$, but not the parameter α . If the tuning parameter gets large, $\lambda \rightarrow \infty$, all strength parameters γ_r are estimated as identical and the total strength is determined solely by $\mathbf{x}_r^T \alpha$ as in the model proposed by Springall (1973). By using a regularization term with positive tuning parameter the parameters are defined and estimable, compare also Friedman et al. (2010), where this procedure has been used in overparameterized multinomial regression models. The choice between fixed and random effects was discussed extensively in Townsend et al. (2013). One advantage of fixed effects models is that they do not have to assume that random effects and covariates are uncorrelated although this is not a problem in the present application. The bigger advantage is that fusion penalties are easily incorporated into the estimation procedure.

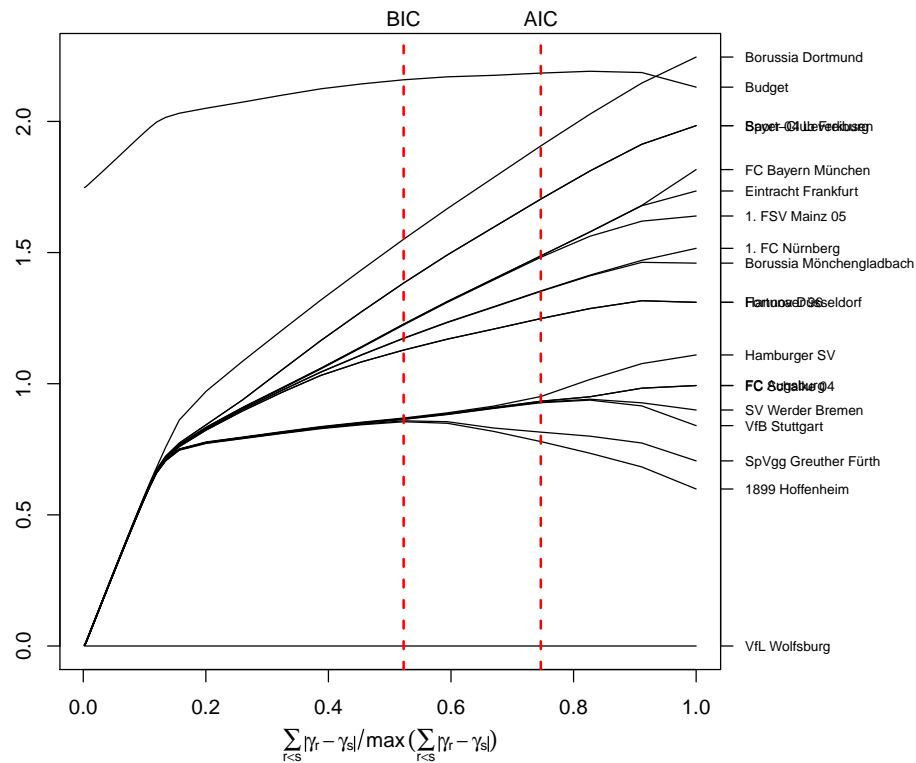


Figure 8.4.: Coefficient paths for ability parameters in the model with a global home advantage and the budget (in 100 millions) using an adaptive L_1 -penalty

In Subsection 8.5.2, we will show that the fixed effects approach with penalties is able to estimate parameters. But first we show how the performance of football teams in the German Bundesliga can be explained by the budget. We use budgets as published by the German sports magazine Kicker (Kicker, August 20, 2012) given in millions. Figure

8.4 shows the coefficient paths for the coefficients plotted against varying strength of the constraints. Here, we use budget in 100 millions for better visibility of the coefficient path. It is seen that the effect of budget is very stable across constraints. As expected, when including the budget different clusters are found because now the γ -parameters represent the abilities that are not explained by the budget. For example, now Borussia Dortmund forms a cluster of its own, whereas Bayern München is in a cluster together with Eintracht Frankfurt and Mainz.

The estimated parameter $\hat{\alpha} = 2.16$, obtained for λ chosen by BIC, implies strong dependence on the budget. In order to get an impression on the reliability of the parameter estimate of the budget at the BIC-optimal λ , we conducted a parametric bootstrap analysis. The corresponding bootstrap confidence interval for $\hat{\alpha}$ is $[1.55; 2.77]$; it supports that budget does have an influence on the team abilities that is not to be neglected.

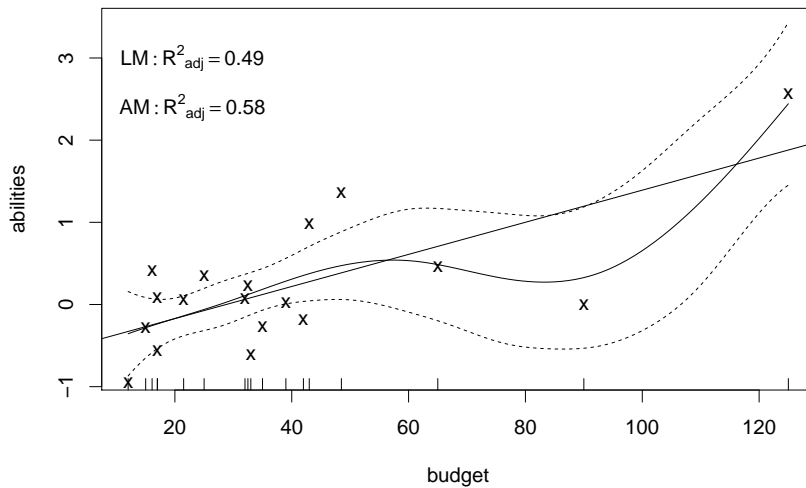


Figure 8.5.: Budgets (in millions) versus estimated abilities for all teams from the Bundesliga season 2012/2013; lines represent linear and additive model fit

The effect of the budget can also be tackled in a different way. In Figure 8.5 the estimated abilities are plotted against the budget. In addition, it shows the fit of a linear regression model and a smoothed version. The smooth model was fitted by use of penalized B-splines (also called P-splines), see Eilers and Marx (1996), with the smoothing parameter chosen by the generalized cross-validation (GCV) criterion. Up to about 70, the linear model fits well, beyond 70 the fit of the non-linear model is determined by just two observations, Wolfsburg and Bayern München. The adjusted R-squared of the linear model is 0.49, that means almost 50% of the variation in abilities is explained by the budget. For the non-linear model the value increases to 0.58 but one can suspect over-fitting. When accepting

the linear model as a simple model that shows almost the same explanatory strength as the non-linear model, one can infer that Wolfsburg (with a budget of 90) is an underachiever. Given the high budget, which is partly due to the fact that the city of Wolfsburg is the home of Volkswagen, the ability is rather low. This holds even in the non-linear model. Bayern München, the club with the highest budget, still shows a positive deviation from the fitted expectation, which is strong for the linear and weak for the the non-linear model. A distinct overachiever is Dortmund (budget of 48.5), which shows one of the strongest deviations from both models. Beyond the identification of over- and underachievers, it is seen that budget is a strong explanatory variable for the ability of a team.

8.5.2. Evaluation of Penalized Estimation

Since parameters are not identifiable maximum likelihood can not be used to estimate the parameters in the model with explanatory variables. We demonstrate in a small simulation study that penalized estimation procedures are able to solve the identifiability problem and can be used to obtain sensible estimates. As true coefficients, we chose values derived from the coefficient estimates of the model fit for the real data from the Bundesliga. We used the thresholds $\theta_1 = -1.66$, $\theta_2 = -0.65$, $\delta = 0.29$ and the budget parameter $\alpha = 2.13$. The team abilities were divided into 5 groups with the coefficients $\gamma_1 = \gamma_2 = \gamma_3 = 2.07$, $\gamma_4 = \gamma_5 = \gamma_6 = 1.73$, $\gamma_7 = \gamma_8 = \gamma_9 = \gamma_{10} = 1.40$, $\gamma_{11} = \gamma_{12} = \gamma_{13} = \gamma_{14} = \gamma_{15} = \gamma_{16} = \gamma_{17} = 0.88$, $\gamma_{18} = 0$.

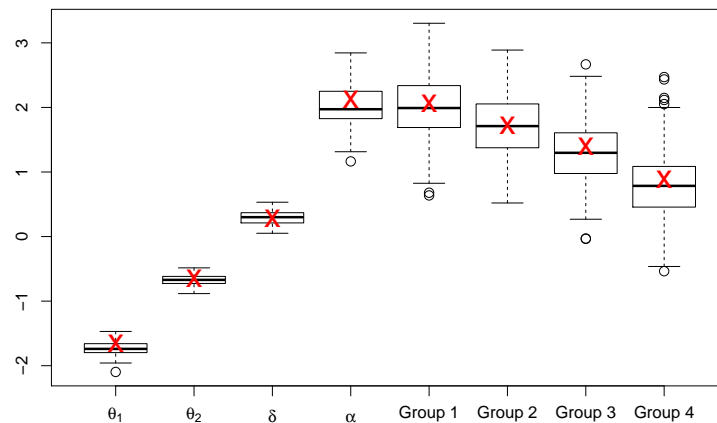


Figure 8.6.: Box plots of coefficient estimates for 100 simulation iterations; estimates for teams with equal abilities are collected in one box; stars denote true values

Figure 8.6 shows the box plots for 100 simulations when using a small tuning parameter λ . Stars denote the true parameter values. In particular, the threshold parameters and the

home advantage parameter are estimated with high accuracy. As expected, the variation of estimates is stronger for the abilities. But, and most important, the parameter of the explanatory variable is estimated rather well. It should be noted that for larger values of the tuning parameter, for example, when λ is chosen by an information criterion, due to shrinkage effects the team effects will be slightly biased.

8.6. Application to the Data from Bundesliga Season 2013/2014

All previous sections in this chapter considered data from the Bundesliga season 2012/2013. In this section, we present all analyses from the previous sections applied to the respective data from the Bundesliga season 2013/2014 and compare the new results to the results from the season 2012/2013.

8.6.1. Ranks and Abilities

	Points	Home	Away	Ability	QSE	Rank
FC Bayern München	90	46	44	2.541	0.366	1
Borussia Dortmund	71	35	36	1.360	0.311	2
FC Schalke 04	64	38	26	1.313	0.310	3
SV Bayer 04 Leverkusen	61	33	28	1.104	0.305	4
VfL Wolfsburg	60	36	24	1.051	0.305	5
Borussia Mönchengladbach	55	36	19	0.927	0.303	6
FSV Mainz 05	53	33	20	0.455	0.300	9
FC Augsburg 1907	52	30	22	0.568	0.300	8
1899 Hoffenheim	44	27	17	0.613	0.300	7
Hannover 96	42	29	13	0.013	0.304	13
Hertha BSC Berlin	41	21	20	0.232	0.301	10
Werder Bremen	39	24	15	0.000	0.304	15
Eintracht Frankfurt	36	20	16	0.122	0.302	11
Sport-Club Freiburg	36	22	14	0.012	0.304	14
VfB Stuttgart	32	19	13	0.092	0.303	12
Hamburger SV	27	18	9	-0.472	0.313	16
1. FC Nürnberg	26	14	12	-0.498	0.314	17
Eintracht Braunschweig	25	18	7	-0.512	0.315	18

Table 8.5.: Final ranking of the German Bundesliga 2013/2014 including points in home matches and away matches; the last three columns show the estimated abilities, quasi standard errors and the ranking corresponding to the estimated abilities for the ordered model including a home advantage parameter

Table 8.5 shows the final ranking of the German Bundesliga in the season 2013/2014. As in the previous season, Bayern München won the championship. The table shows the points

each team won in total and separated between points won at home and away. The abilities are estimated by an (unpenalized) model with a global home advantage. It can be seen that the abilities would result in a slightly different ranking than the points. For example, according to the estimated abilities Stuttgart would perform much better than according to their points, for Werder Bremen we see the contrary effect.

8.6.2. Team-specific Home Effects

For the abilities in Table 8.6, a model with team-specific home advantages (or home effects, few teams actually have a home disadvantage) is fitted. For example, in away matches Hertha BSC Berlin has a considerably higher ability than in home matches. Nevertheless, even Hertha BSC Berlin won one point more in home matches than in away matches (see Table 8.5).

	Overall		Home		Away	
	Ability	Rank	Ability	Rank	Ability	Rank
FC Bayern München	2.541	1	2.522	1	2.968	1
Borussia Dortmund	1.360	2	1.248	7	1.871	2
FC Schalke 04	1.313	3	2.171	2	0.965	4
SV Bayer 04 Leverkusen	1.104	4	1.439	5	1.179	3
VfL Wolfsburg	1.051	5	1.734	4	0.809	5
Borussia Mönchengladbach	0.927	6	1.814	3	0.384	9
FSV Mainz 05	0.455	9	1.253	6	-0.112	14
FC Augsburg 1907	0.568	8	0.836	9	0.556	7
1899 Hoffenheim	0.613	7	1.189	8	0.416	8
Hannover 96	0.013	13	0.815	10	-0.525	17
Hertha BSC Berlin	0.232	10	0.140	16	0.637	6
Werder Bremen	0.000	15	0.260	14	0.000	13
Eintracht Frankfurt	0.122	11	0.396	11	0.235	10
Sport-Club Freiburg	0.012	14	0.292	13	0.088	12
VfB Stuttgart	0.092	12	0.393	12	0.172	11
Hamburger SV	-0.472	16	-0.332	17	-0.332	16
1. FC Nürnberg	-0.498	17	-0.571	18	-0.198	15
Eintracht Braunschweig	-0.512	18	0.174	15	-1.023	18

Table 8.6.: Comparison of the estimated abilities from the model with a global home advantage to the estimated abilities from the model with team-specific home advantages

Borussia Dortmund seems to show the most prominent difference between the abilities on home ground and away. While Dortmund is the second best team away it is only the 7th best team on home ground. This is even more noteworthy as the stadium of Dortmund is often considered to be the best football stadium in the world with the highest capacity in the Bundesliga. Therefore, the support of Dortmund in its own stadium is extraordinary, after all the performance of Dortmund is better in away matches than in home matches.

8.6.3. Identification of Clusters

Figure 8.7 shows the coefficient paths for the abilities of the single teams in the model with a global home advantage. The abilities (or rather the differences between the abilities) are penalized with an L_1 -penalty and, therefore, shrunk toward zero. This enforces clusters within the teams where a cluster entails teams with similar abilities.

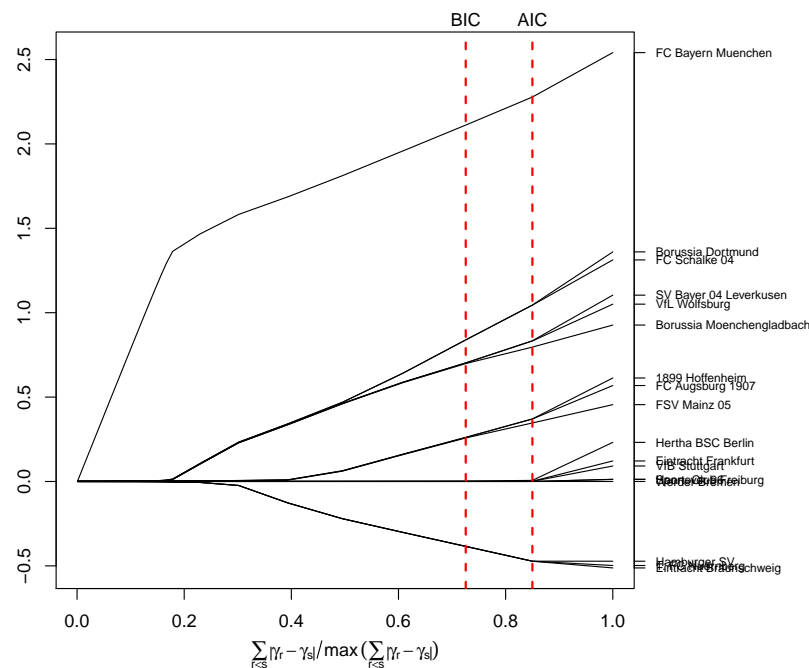


Figure 8.7.: Coefficient paths for ability parameters in the model with a global home advantage using an adaptive L_1 -penalty.

Cluster	Ability
1 FC Bayern München	2.11
2 Borussia Dortmund; FC Schalke 04	0.84
3 Borussia Mönchengladbach; Bayer 04 Leverkusen; VfL Wolfsburg	0.70
4 1. FSV Mainz; FC Augsburg; 1899 Hoffenheim;	0.26
5 Sport-Club Freiburg; Eintracht Frankfurt; Hannover 96; Hertha BSC Berlin VfB Stuttgart; SV Werder Bremen	0.00
6 1. FC Nürnberg; Eintracht Braunschweig; Hamburger SV	-0.38

Table 8.7.: Clusters of teams with corresponding abilities.

Table 8.7 shows the resulting clusters if the optimal path point in Figure 8.7 is chosen by BIC. In total we end up with 6 clusters while there were 8 clusters in table 8.3 for the season 2012/2013. Like in the previous season, the first cluster is the champion Bayern München, seemingly playing in a league of its own. The last cluster entails the two relegated teams

Braunschweig and Nürnberg as well as the Hamburger SV, who had to play relegation matches to stay up in the Bundesliga. In the previous season, the last three teams had been split into two clusters.

8.6.4. Clustering of Teams and Home Effects

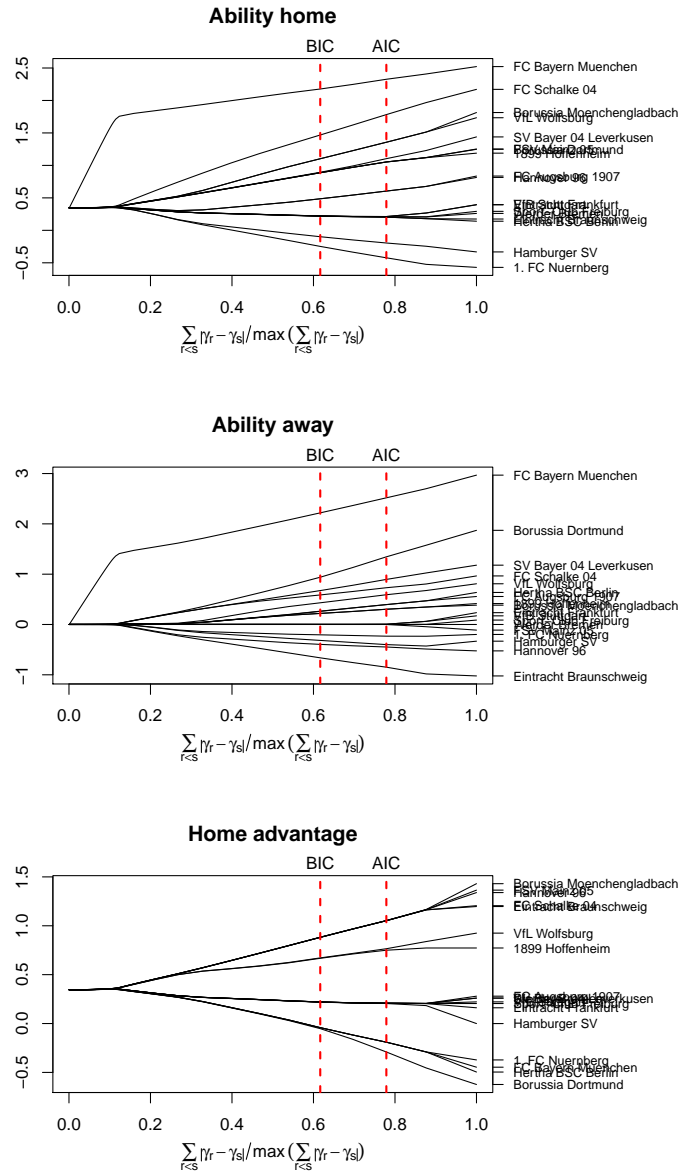


Figure 8.8.: Coefficient paths for home abilities, away abilities and home advantages in the model with team specific home advantages using an adaptive L_1 -penalty

Figure 8.8 shows the paths for the model with team-specific home effects, compare Figure 8.3 for season 2012/2013. Again, it can be seen that Borussia Dortmund has the most negative home effect (for the unpenalized case). While the abilities home and away show

more different clusters than in the season 2012/2013, the home advantage only has four different clusters if model selection is done by BIC.

8.6.5. Accounting for Explanatory Variables

In this subsection, the inclusion of the budget of the teams is considered. In Figure 8.9,

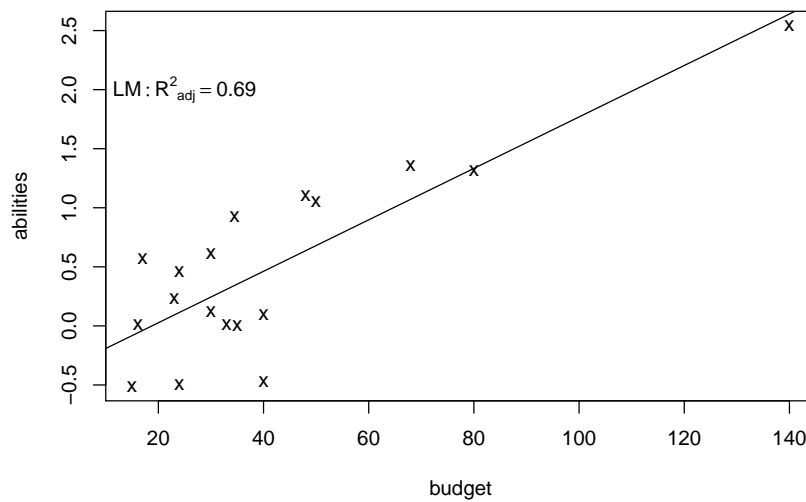


Figure 8.9.: Budgets (in millions) versus estimated abilities for all teams from the Bundesliga season 2013/2014

the budget (in millions) of the teams is plotted against the respective estimated abilities. The solid line represents the LS estimator of the respective linear model ($R^2 = 0.69$). In contrast to the data from 2012/2013, the correlation is clearly linear, the fit of an additive model resulted in a linear model. The plot shows that the abilities of a team highly depend on the budget of the respective club.

Figure 8.10 shows the coefficient paths for the ability parameters if the budget is incorporated in the model (and a global home effect). Again, this results in clusters of teams with similar abilities. In this case, the abilities are interpreted as the abilities if the budget is already eliminated by the model. Again, if the optimal path point is chosen by BIC 6 clusters are found. Taking the budget into account, the Hamburger SV had the worst performance of all teams. The best performance had the cluster of the teams Augsburg, Mönchengladbach, Leverkusen, Wolfsburg, Hoffenheim, Mainz and Dortmund. Bayern München, although having a season with total dominance of the league, only appear in the second cluster. The effect strength of the budget is very similar to the effect in the previous season.

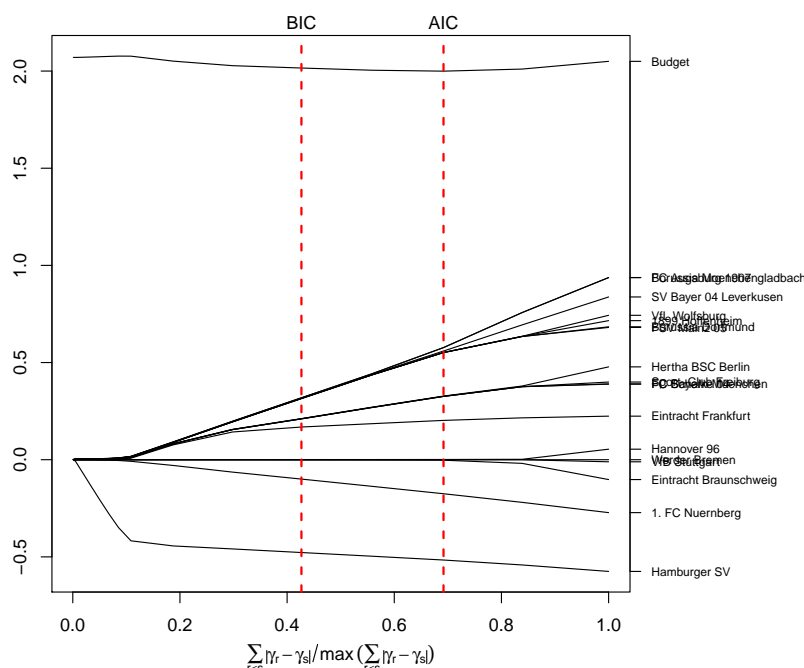


Figure 8.10.: Coefficient paths for ability parameters in the model with a global home advantage and the budget (in 100 millions) using an adaptive L_1 -penalty

8.7. Concluding Remarks

All calculations in this chapter have been conducted by using the statistical software **R** (R Core Team, 2015). Most of the available add-on packages for paired comparison models in **R** are restricted to the case of binary response and cannot deal with ordered response. The most popular packages are **prefmod** (Hatzinger and Dittrich, 2012) and **BradleyTerry2** (Turner and Firth, 2012). The former uses the log linear representation of BT-models and can handle draws in the response variable. The latter can also handle covariates by assuming random effects for the ability parameters but only in the case of binary responses.

Here we favor a direct approach to the fitting of ordinal paired comparison models (without regularization) that is based on the embedding into the framework of generalized linear models. By including the restrictions on the thresholds and the construction of specific design matrices that include the effect of home advantages BT models for ordered response can be fitted by using the add-on package **VGAM** (Yee, 2010). It also allows to use alternative link functions. The procedure, but without team-specific covariates and regularization, has been implemented in the package **ordBTL** (Casalicchio, 2013).

In our extended framework we have to also include penalty terms. A very general approach that allows to combine a variety of different penalties in univariate GLMs has been proposed by Oelker and Tutz (2015), and is available in the package `gvcm.cat` (Oelker, 2015). With the help of Margret Oelker it has been adapted such that also cumulative logit models can be fitted.

9. Prediction of Soccer Tournaments Based on Regularized Poisson Regression

9.1. Introduction

In the last few years various approaches to the statistical modeling of major international soccer events have been proposed, among them the Union of European Football Associations (UEFA) Champions League (CL; Karlis and Ntzoufras, 2011, Eugster et al., 2011), the European football championship (EURO; Leitner et al., 2010a, Zeileis et al., 2012, Groll and Abedieh, 2013) or the Fédération Internationale de Football Association (FIFA) World Cup (Leitner et al., 2010b; Stoy et al., 2010; Dyte and Clarke, 2000). In particular, the current FIFA World Cup 2014 in Brazil is accompanied by various publications trying to predict the tournament winner, see, e.g., Zeileis et al. (2014), Goldman-Sachs Global Investment Research (2014), Silver (2014) and Lloyd's (2014).

In general, statistical approaches to the modeling of soccer data can be divided into two major categories: the first ones are based on the easily available source of "prospective" information contained in bookmakers' odds, compare Leitner et al. (2010a) and their follow-up papers. They already correctly predicted the finals of the EURO 2008 as well as Spain as the 2010 FIFA World Champion and as the 2012 EURO Champion. The winning probabilities for each team were obtained simply by aggregating winning odds from several online bookmakers. Based on these winning probabilities, by inverse tournament simulation team-specific abilities can be computed by paired comparison models. Using this technique the effects of the tournament draw are stripped. Next, pairwise probabilities for each possible game at the corresponding tournament can be predicted and, finally, the whole tournament can be simulated. Using this approach, Zeileis et al. (2014) predicted the host Brazil to win the FIFA World Cup 2014 with a probability of 22.5%, followed by Argentina (15.8%) and Germany (13.4%).

This chapter is a modified version of Groll et al. (2015), previous work on the issue can be found in the technical report 166 (Groll et al., 2014). See Chapter 1 for more information on the personal contributions of all authors and textual matches.

It should be noted that this method will always predict the team that has the lowest (average) bookmaker odds as the tournament winner and, hence, is implicitly assuming that all available information is covered by the bookmakers expertise. This is not unrealistic, as one can indeed expect bookmakers to use sophisticated models when setting up their odds, as they have strong economic incentives to rate the team strengths of soccer teams correctly. Although the bookmakers' models certainly contain covariate information of the competing teams, at least indirectly, an alternative approach is to explicitly model the influence of covariates on the success of soccer teams.

This task leads to the second category of approaches that are based on regression models. A simple standard linear regression approach was used by Stoy et al. (2010) to analyze the success of national teams at FIFA World Cups. The success of a team at a World Cup is measured by a defined point scale that is supposed to be normally distributed. Beside some sport-specific covariates also political-economic, socio-geographic as well as some religious and psychological influence variables are considered. Based on this model, a prediction for the FIFA World Cup 2010 was obtained.

In contrast to Stoy et al. (2010), most of the regression approaches directly model the number of goals scored in single soccer matches, assuming that the number of goals scored by each team follows a Poisson distribution model, see, e.g., Maher (1982), Lee (1997), Dixon and Coles (1997), Dyte and Clarke (2000), Rue and Salvesen (2000) and Karlis and Ntzoufras (2003). For example, Dyte and Clarke (2000) predict the distribution of scores in international soccer matches, treating each team's goals scored as conditionally independent Poisson variables depending on two influence variables, the FIFA world ranking of each team and the match venue. Poisson regression is used to estimate parameters for the model and based on these parameters the matches played during the 1998 FIFA World Cup can be simulated.

Similarly, Goldman-Sachs Global Investment Research (2014) set up a regression model based on the entire history of mandatory international football matches—i.e., no friendlies—since 1960, ending up with about 14,000 observations. The dependent variable is the number of goals scored by each side in each match, assuming that the number of goals scored by a particular side in a particular match follows a Poisson distribution. They incorporate six explanatory covariates: the difference in the Elo rankings¹ between the two teams, the average number of goals scored/received by the competing teams over the last ten/five mandatory international games, a dummy variable indicating whether the regarding match was a World Cup match, a dummy variable indicating whether the considered team played in its home country, a team-specific dummy variable indicating whether the considered team played on its home continent. Finally, based on the estimated regression parameters, a probability

¹ The Elo ranking is a composite measure of national football teams' success, which is based on the entire historical track record and which, in contrast to the FIFA ranking, is available for the entire history of international football matches (see Elo, 2008).

distribution for the outcome of each game is obtained and a Monte Carlo simulation with 100,000 draws is used to generate the probabilities of teams reaching particular stages of the tournament, up to winning the championship. The forecast tournament winner at the FIFA World Cup 2014 is Brazil with a rather high winning probability of 48.5%, followed by Argentina (14.1%) and Germany (11.4%).

At this point, we also want to mention other prediction approaches, which cannot be classified into one of the two aforementioned major categories of statistical approaches for modeling soccer data. For example, Dobson and Goddard (2011) or Forrest and Simmons (2000) use discrete choice models for the modeling of match outcomes. Concerning the prediction of the FIFA World Cup 2014, an approach proposed by Silver (2014) is based on the so-called Soccer Power Index (SPI). The SPI is a rating system, which uses historical data on both the international and club level to predict the outcome of a match. The algorithm uses several years of data, taking into account goals scored and allowed, quality of the lineup fielded, and the location of the match. In addition, the index weights recent matches more heavily, and also takes into account the importance of the match – e.g., World Cup matches count much more than friendly matches. Based on the SPI, Silver (2014) forecasts again Brazil as the tournament winner at the FIFA World Cup 2014, also with a rather large winning probability of 45.2%, followed by Argentina (12.8%) and Germany (11.9%).

The other alternative approach is from a more economical perspective: the London insurance market Lloyd's (2014) uses players wages and endorsement incomes together with a collection of additional indicators to construct an economic model, which estimates players incomes until retirement. These projections form the basis for assessing insurable values by players age, playing position and nationality. As Germany and Spain are associated with the largest estimated insured values, according to this approach they turn out to be the top favorites for winning the current World Cup.

In the approach that we propose here we focus on international soccer tournaments, here applied to FIFA World Cups, and use a Poisson model for the number of goals scored by competing teams in the single matches of the tournaments. Several potential influence variables are considered and, additionally, team-specific effects are included in the form of fixed effects, resulting in a flexible generalized linear model (GLM). Incorporating a method for the selection of relevant predictors, we obtain a regularized solution for our model. The variable selection is based on suitable L_1 -penalization techniques and is performed with the `grplasso` function from the corresponding R-package (see Meier et al., 2008). As an application, the approach is used to fit data from previous FIFA World Cups and finally, based on the obtained estimates, the FIFA World Cup 2014 is predicted.

It should be noted that in contrast to other team sports, such as basketball, ice-hockey or handball, in soccer pure chance plays an important role. A major reason for this is that, compared to other sports, in soccer fewer points (goals) are scored and thus singular game

situations can have a tremendous effect on the outcome of the match. One consequence is that every now and then alleged (and unpredictable) underdogs win tournaments. There are countless examples in history for such events, throughout all competitions. We want to mention only some of the most famous ones: Germany's first World Cup success in Switzerland 1954, known as the "miracle from Bern"; Greece's victory at the EURO 2004; FC Porto's triumph in the UEFA CL season 2003/2004. Nevertheless, it can be interesting to investigate the relationship and dependency structure between different potentially influential covariates and the success of soccer teams (in our case in terms of the number of goals they score).

The rest of the chapter is structured as follows. In Section 9.2, we introduce the team-specific Poisson model for the number of goals. Section 9.3 entails a description of the data for the application to the FIFA World Cup, including a list of possible influence variables. Furthermore, the model is fitted to the data and used to predict the FIFA World Cup 2014. Note that all computations have been performed by use of the statistical software R (R Core Team, 2015).

9.2. Model and Estimation

Our proposed model concentrates on the number of goals a team scores against a specific opponent. For every team, specific attack and defense parameters are considered. Furthermore, the covariates of both teams, which might have an influence on the number of scored goals, are considered in the form of differences.

Let for n teams $y_{ijk}, i, j \in \{1, \dots, n\}, i \neq j$, denote the number of goals scored by team i when playing team j at tournament k . The considered model has the form:

$$y_{ijk} | \mathbf{x}_{ik}, \mathbf{x}_{jk} \sim \text{Pois}(\lambda_{ijk})$$

$$\log(\lambda_{ijk}) = \beta_0 + (\mathbf{x}_{ik} - \mathbf{x}_{jk})^\top \boldsymbol{\beta} + \text{att}_i - \text{def}_j. \quad (9.1)$$

It is assumed that the number of goals that team i scores follows a Poisson distribution with given team-specific parameters and covariates of both teams. In addition, the two observations of one match are assumed to be independent, given the team-specific parameters and covariates.

The linear predictor consists of the attacking parameter att_i of the team i and the defending parameter def_j of its opponent j . The covariates of team i at tournament k are collected in a vector $\mathbf{x}_{ik} = (x_{ik1}, \dots, x_{ikp})^\top$ of length p . In the following, we assume that the covariates of each team can vary over different tournaments (but not within a tournament). Each

covariate is incorporated as the difference between the respective covariates of both teams. The covariate effects are collected in the vector $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top$ and β_0 represents the intercept.

If the linear predictor of the model is re-formulated, it can be denoted by

$$\begin{aligned}\eta_{ijk} &= \beta_0 + (\mathbf{x}_{ik} - \mathbf{x}_{jk})^\top \boldsymbol{\beta} + att_i - def_j \\ &= \beta_0 + \mathbf{x}_{ik}^\top \boldsymbol{\beta} + att_i - \mathbf{x}_{jk}^\top \boldsymbol{\beta} + def_j \\ &= \beta_0 + \gamma_i - \delta_j.\end{aligned}$$

Here, $\gamma_i = \mathbf{x}_{ik}^\top \boldsymbol{\beta} + att_i$ and $\delta_j = \mathbf{x}_{jk}^\top \boldsymbol{\beta} + def_j$ represent the total attack ability of team i and defense ability of team j , respectively. Hence, att_i and def_i act as additional parameters covering ability differences that are not covered by the covariate effects yet. This denotation emphasizes that the model can be seen as a paired comparison model as the linear predictor is mainly composed of the difference of the abilities of two opponents.

Generally, the estimation of the covariate effects will be obtained by regularized estimation approaches. The idea is to first set up a model with a rather large number of possibly influential variables and then to regularize the effect of the single covariates. This regularization aims at diminishing the variance of the parameter estimates and, hence, to provide lower prediction error than the unregularized maximum likelihood estimator. The basic concept of regularization is to maximize a penalized version of the log-likelihood $l(\boldsymbol{\alpha})$ where $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_p)^\top$ represents a general parameter vector. More precisely, one maximizes the penalized log-likelihood

$$l_p(\boldsymbol{\alpha}) = l(\boldsymbol{\alpha}) - \lambda J(\boldsymbol{\alpha}), \quad (9.2)$$

where λ represents a tuning parameter, which is used to control the strength of the penalization. In practice, this tuning parameter has to be chosen either by suitable criteria for model selection or by cross-validation. Model selection criteria are usually based on a compromise between the model fit (e.g. in terms of the likelihood) and the complexity of the model, like AIC (Akaike, 1973) or BIC (Schwarz, 1978). The penalty term $J(\boldsymbol{\alpha})$ can have many different shapes. Hoerl and Kennard (1970) suggested the so-called ridge penalty

$$J(\boldsymbol{\alpha}) = \sum_{i=1}^p \alpha_i^2,$$

where the sum over the squares of all parameters in the model is penalized. The ridge penalty has the feature to shrink the respective parameter estimates towards zero. After all,

ridge cannot set estimates to zero exactly and, therefore, can not perform variable selection. In our analysis, we will use a penalty based on the absolute values of the parameters instead of the squared values resulting in a so-called least absolute shrinkage and selection operator (LASSO) penalty. The LASSO estimator was originally proposed by Tibshirani (1996) and uses the penalty

$$J(\boldsymbol{\alpha}) = \sum_{i=1}^p |\alpha_i|. \quad (9.3)$$

In contrast to the ridge penalty, LASSO can provide parameter estimates, which are exactly zero and, therefore, enforces variable selection.

The team-specific ability parameters att_i and def_j are considered as fixed effects and are coded by dummy variables within the design matrix. From this perspective, the attack (and, analogously, the defense) variables are seen as categorical covariates with as many categories as there are teams². One assigns 1 to the dummy variables associated with att_i , if the goals of team i are considered, and 0 otherwise. Similarly, one assigns -1 to the dummy variables associated with def_j , if team j is the opponent, and 0 otherwise. An extract of the corresponding design matrix is given in Table 9.2.

In the following, both team-specific effects corresponding to one team are treated as a group. Hence, the original LASSO penalty from equation (9.3) has to be modified appropriately according to the so-called Group LASSO penalty proposed by Yuan and Lin (2006). The Group LASSO penalizes the L_2 -norm of the respective parameter vectors $(att_1, def_1)^T, \dots, (att_n, def_n)^T$. Hence, the parameters of attack and defense abilities of single teams are simultaneously shrunk towards zero and, if shrunk exactly to zero, excluded from the model. Besides, the covariate effects $\boldsymbol{\beta}$ are penalized using the ordinary LASSO penalty from equation (9.3). Altogether, the penalty term for model (9.1) is given by

$$J(\boldsymbol{\alpha}) = \sum_{i=1}^p |\beta_i| + \sqrt{2} \sum_{i=1}^n \sqrt{att_i^2 + def_i^2}.$$

The prefactor $\sqrt{2}$ controls for the group sizes of the groups of team-specific parameters, compare Yuan and Lin (2006) or Meier et al. (2008). Another advantage of penalization is the way correlated predictors are treated. If two predictors are highly correlated, the parameter estimates are stabilized by the penalization. The chosen LASSO penalty tends to include only one of the predictors and only includes the second predictor if it entails additional information for the response variable. Therefore, if several variables possibly contain information on the strength of teams they can be used simultaneously. The most

² Usually, for reasons of identifiability, categorical predictors with k factor levels are coded by $k-1$ dummies. However, the regularization approach introduced in the following (with $\lambda > 0$) provides unique estimates despite the issues of identifiability.

informative variable is chosen automatically by the penalty term. The model can easily be fitted by use of the `grplasso` function from the corresponding R-package (see Meier et al., 2008).

Note that, alternatively, similar to the model used in Groll and Abedieh (2013) the team-specific effects could be estimated as random effects. Then, the attack and the defense parameter of team i are assumed to be multivariate normally distributed. In this case, the ability parameters are automatically regularized by the assumption of a distribution and only the covariate effects β are explicitly penalized by using LASSO. The algorithm `glmLasso` proposed in Groll and Tutz (2014) can be used to fit this model. However, this results in a model more focused on team-specific effects than covariate effects due to the different, namely lower, penalization of the random team-specific effects. Therefore, this modeling approach is not pursued in the following.

9.3. Application

In the following, the proposed model is applied to data from previous FIFA World Cups and is then used to predict the FIFA World Cup 2014 in Brazil.

9.3.1. Data

In this section, we give a brief description of the used covariates. For each participating team, the covariates are observed either for the year of the respective World Cup (e.g. GDP per capita) or shortly before the start of the World Cup (e.g. FIFA ranking). Therefore, the covariates of a team vary from one World Cup to another and, hence, the model allows for a prediction of a new World Cup based on the current covariate realizations.

Economic factors

GDP per capita. The gross domestic product (GDP) per capita represents the economic strength of a country. To account for the general increase of the GDP, a ratio of the GDP per capita of the respective country and the worldwide average GDP per capita is used. The GDP data were collected from the website of the United Nations Statistics Division (<http://unstats.un.org/unsd/snaama/dnllist.asp>).

Population. The population size of a country may have an influence on the success of a national team as small countries will have a smaller amount of players to choose from. The population size is used as a ratio with the respective global population to account for

the general growth of the world population. The data source is the website of the world bank (<http://data.worldbank.org/indicator/SP.POP.TOTL>).

Sportive factors

ODDSET odds. Bookmakers' odds on the probability to win a World Cup already entail a great amount of covariates and information about the respective team and, therefore, can be assumed to be a good predictor for the success of a national team? The odds were provided by the German state betting agency ODDSET. The bookmakers' odds are converted into winning probabilities by taking the inverse of the odds followed by elimination of the bookmakers' margin. Hence, the variable reflects the probabilities of ODDSET for each team to win the respective World Cup³.

FIFA ranking. The FIFA ranking provides a ranking system for all national teams measuring the performance of the team over the last four years. The exact formula for the calculation of the FIFA points and all rankings since implementation of the FIFA ranking system can be found at the official FIFA website (<http://de.fifa.com/worldranking/index.html>). Since the calculation formula of the FIFA points changed after the World Cup 2006, the rankings according to FIFA points are used instead of the points⁴.

Home advantage

Host. The host of the World Cup could have an advantage over its opponents because of the stronger support of the crowd in the stadium. Therefore, a dummy variable for the respective host of the World Cup is included.

Continent. Before the World Cup 2014, many discussions revolved around the climatic conditions in Brazil and who would deal best with these conditions. One could assume that teams from the same continent as the host of the World Cup (including the host itself) may have advantages over teams from other continents, as they might better get along with the climatic and cultural circumstances. A dummy variable for the continent of the World Cup host is included.

³ The possibility of betting on the overall cup winner before the start of the tournament is quite novel. For example, the German state betting agency ODDSET offered the bet for the first time at the FIFA World Cup 2002.

⁴ The FIFA ranking was introduced in August 1993.

Factors describing the team's structure

The following variables are thought to describe the structure of the teams. Each variable was observed with the final squad of 23 players nominated for the respective World Cup.

(Second) maximum number of teammates. If many players from one club play together in a national team, this could lead to an improved performance of the team as the teammates know each other better. Therefore, both the maximum and the second maximum number of teammates from the same club are counted and included as covariates.

Average age. The average age of all 23 players is observed to include possible differences between rather old and rather young teams.

Number of Champions League (Europa League) players. The European club leagues are valued to be the best leagues in the world. Therefore, the competitions from teams between the best European teams, namely the UEFA Champions League and the UEFA Europa League (previously UEFA Cup) can be seen as the most prestigious and valuable competitions on club level. As a measurement of the success of the players on club level, the number of players in the semi finals (taking place only weeks before the respective World Cup) of these competitions are counted.

Number of players abroad. Finally, the national teams strongly differ in the numbers of players playing in a league of the respective country and players from leagues of other countries. For each team, the number of players playing in clubs abroad (in the season previous to the respective World Cup) are counted.

Factors describing the team's coach

Also covariates of the coach of the national team may have an influence on the performance of the team. Therefore, the *age* of the coach and the duration of the *tenure* of the coach are observed. Furthermore, a dummy variable is included, if the coach has the same *nationality* as his team or not.

Unfortunately, the covariate *ODDSET odds* is not available before the FIFA World Cup 2002. But as this covariate can be assumed to contain already a lot of expertise and information about an upcoming World Cup, we decided to perform a separate analysis for the FIFA World Cup data from 2002-2010 (from now on denoted by WC2002), including the odds. But as this results in a quite small data basis, another analysis will be performed on a data set including the World Cups from 1994-2010, excluding the covariate *ODDSET odds* (from now on denoted by WC1994).

Note that the differences of the three binary variables *host*, *continent* and *nationality*, which originally have been encoded with $\{0, 1\}$, lead to new categorical variables with the three factor levels -1, 0 and +1. For each of these new categorical covariates we use dummy encoding with -1 as the reference category and, hence, obtain two columns per covariate in the design matrix, e.g. *host0* and *host1*, corresponding to the factor levels 0 and 1, respectively. The dummy variables corresponding to one categorical covariate are treated as groups and, hence, are also penalized by a Group LASSO penalty, similar to the attack and defense ability parameters.

It should be noted that at the FIFA World Cup 2014 the national team of Bosnia and Herzegovina participated for the very first time. Therefore, for this team no estimates of its team-specific effects are available. Analogously, the national team of Colombia participating also at the FIFA World Cup 2014 did not participate in any of the FIFA World Cups from 2002-2010. In order to obtain nonetheless reasonable estimates for the team-specific effects of such teams, which can then be used for the prediction of the FIFA World Cup 2014, we collect all teams that have only participated once in the tournaments from the respective data basis in a group called *newcomers*. Therefore, these teams share the same team-specific ability parameters. Exemplarily, for the WC2002 data this concerns the following 12 teams: Angola, China, Czech Republic, Ireland, New Zealand, North Korea, Senegal, Slovakia, Togo, Trinidad & Tobago, Turkey, Ukraine.

As already mentioned, in the model specification of model (9.1) from Section 9.2 all covariates are considered in the form of differences. For example, in the first match of the FIFA World Cup 2002 in Japan and South Korea, where France played against Senegal (which is among the group of *newcomers* in our sample), the French team had an *average age* of 28.30 years, was on first place in the current *FIFA ranking* and had a winning probability given by the *ODDSET odds* of 15%, while Senegal's team had an *average age* of 24.30 years, was on 42th place in the current *FIFA ranking* and had a winning probability of 1%. Hence, when the goals of France are considered, this results in the following differences for the metric covariates: $age = 28.30 - 24.30 = 4.00$, $rank = 1 - 42 = -41$, $odds = 0.15 - 0.01 = 0.14$. For the categorical variable $host \in \{-1, 0, 1\}$ we get $host = 0 - 0 = 0$, which results in the entries $host0 = 1$ and $host1 = 0$ in the two columns of the design matrix corresponding to the dummy encoding, as the factor level -1 was chosen as the reference category. An extract of the design matrix part, which corresponds to the covariates is presented in Table 9.1. The matrix resulting from the encoding of the team-specific effects is illustrated in Table 9.2.

Goals	Team	Opponent	Age	Rank	Odds	Host0	Host1	...
0	France	Newcomer	4.00	-41	0.14	1	0	...
1	Newcomer	France	-4.00	41	-0.14	1	0	...
1	Uruguay	Denmark	-2.10	4	-0.00	1	0	...
2	Denmark	Uruguay	2.10	-4	0.00	1	0	...
1	Denmark	Newcomer	3.10	-22	0.01	1	0	...
1	Newcomer	Denmark	-3.10	22	-0.01	1	0	...
0	France	Uruguay	3.00	-23	0.14	1	0	...
0	Uruguay	France	-3.00	23	-0.14	1	0	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 9.1.: Extract of the design matrix part which corresponds to the covariates.

FRA.att	FRA.def	NEW.att	NEW.def	URU.att	URU.def	DEN.att	DEN.def
1	0	0	-1	0	0	0	0
0	-1	1	0	0	0	0	0
0	0	0	0	1	0	0	-1
0	0	0	0	0	-1	1	0
0	0	0	-1	0	0	1	0
0	0	1	0	0	0	0	-1
1	0	0	0	0	-1	0	0
0	-1	0	0	1	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 9.2.: Encoding of the team specific-effects

9.3.2. Estimation Results

In this section, we present the fit of model (9.1) from Section 9.2 on the basis of both data sets, i.e. the FIFA World Cups 1994-2010 and 2002-2010, which is then used for the prediction of the FIFA World Cup 2014.

As pointed out in Section 9.2 we use LASSO-type penalization approaches to fit the model (9.1). The crucial step is now to determine the optimal value of the tuning parameter λ from equation (9.2). Note that different levels of sparseness are obtained depending on the selection of the optimal tuning parameter λ . In general, information criteria such as Akaike's information criterion (AIC, see Akaike, 1973) or the Bayesian information criterion (BIC, see Schwarz, 1978), also known as Schwarz's information criterion, could be used, but as our main focus is on achieving good prediction results in order to be able to provide a realistic forecast of the FIFA World Cup 2014, we decided to use 10-fold cross validation (CV) based on the conventional Poisson deviance score ⁵. The corresponding 10-fold CV results are illustrated in Figure 9.1, exemplarily for the WC1994 data.

⁵ As two observations corresponding to the goals of the same match belong together, we do not exclude single observations from the training data, but single matches.

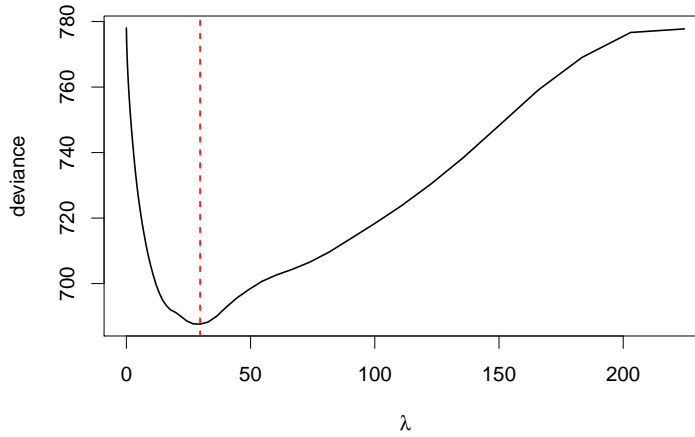


Figure 9.1.: Deviance for 10-fold CV for Model (9.1), exemplarily for the FIFA World Cup data 1994-2010; the optimal value of the penalty parameter λ is shown by the vertical line.

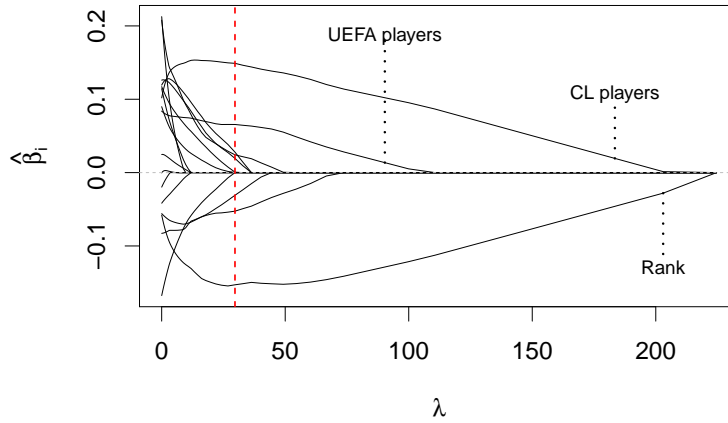
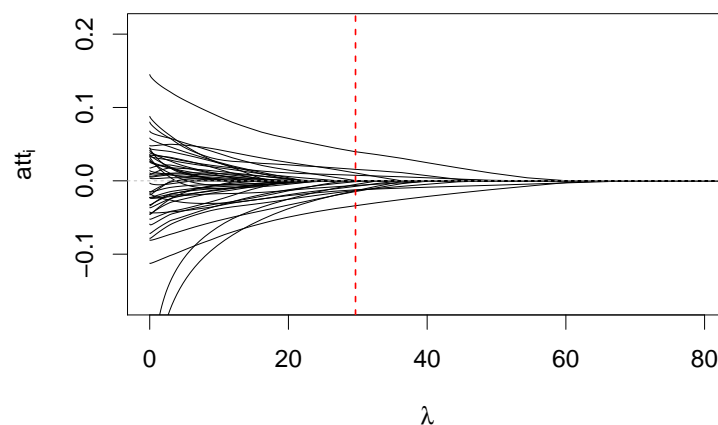
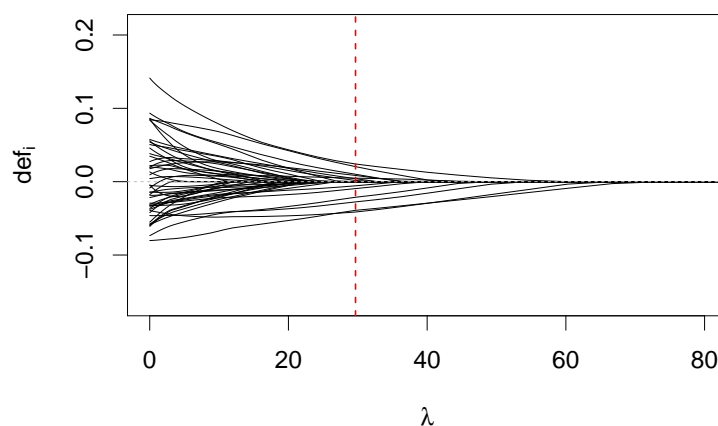


Figure 9.2.: Coefficient paths of the covariate effects vs. the penalty parameter λ , exemplarily for the FIFA World Cup data 1994-2010; the optimal value of the penalty parameter λ is shown by the vertical line.

Additionally, in Figure 9.2 the coefficient paths for the (scaled) covariates are shown along the penalty parameter λ . Note that in order to correctly apply the LASSO algorithms, all covariates (both binary and continuous) were scaled to have mean 0 and variance 1. Besides, Figure 9.3 illustrates the coefficient paths of the team-specific attack and defense parameters. In Table 9.3, the fixed effects estimates for the (scaled) covariates are shown for both data sets. The optimal tuning parameter λ , which minimizes the deviance shown in Figure 9.1, leads to a model with 10 (out of possibly 17) regression coefficients different from zero for the WC1994 data set. The paths illustrated in Figure 9.2 show that the first covariate to be selected is the *FIFA rank*, followed by the *number of CL players* and *number of UEFA players* (when the penalty parameter λ decreases). Together with the fact that the estimated effects of these three covariates also exhibit the highest absolute



(a) Team-specific attack parameters



(b) Team-specific defense parameters

Figure 9.3.: Coefficient paths of the team-specific attack (a) and defense effects (b) vs. the penalty parameter λ , exemplarily for the FIFA World Cup data 1994-2010; the optimal value of the penalty parameter λ is shown by the vertical lines.

values, this indicates that the three covariates offer the highest explanatory power among all regarded covariates. The estimated coefficients show the intuitively expected effects: better, i.e. lower, FIFA ranks and more players that have been successful with their clubs in the UEFA Champions or Europa League have positive effects on the number of goals scored. It is also worth mentioning that at the optimal tuning parameter, for several teams the ability estimates are still zero, compare Figure 9.3.

In general, similar graphs are obtained for the smaller WC2002 data, which includes the *ODDSET odds* as a covariate. The major difference is that the *ODDSET odds* are the first variable to enter the model, followed by the *FIFA rank*. This confirms the supposition

	WC 1994-2010	WC 2002-2010
CL players	0.149	0.075
UEFA players	0.066	0
Age Coach	0	-0.017
Tenure Coach	0	-0.071
Legionaires	0	0
Max. # teammates	0	0
Sec. max # teammates	-0.053	0
Age	0	0
Rank	-0.153	-0.167
GDP	0.024	0.042
Odds	-	0.113
Population	-0.031	-0.060
Continent0	0.001	0.010
Continent1	0.000	-0.003
Nation Coach0	0	0
Nation Coach1	0	0
Host0	0.019	0
Host1	0.028	0

Table 9.3.: Estimates of the covariate effects for the FIFA World Cups 1994-2010 and 2002-2010.

that the bookmakers' odds cover already a lot of information and, hence, provide strong explanatory power in the context of the success of soccer teams. Again, also the *number of CL players*, the third covariate that enters the model, seems to play an important role.

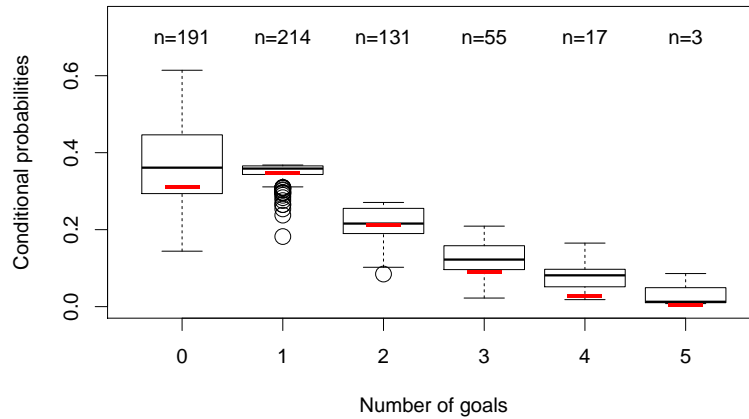
The model including the odds is sparser with only 9 out of 18 regression coefficients different from zero. A possible explanation is that the *ODDSET odds* already include a lot of information from other covariates, as for example the host effect, which has been found in the WC1994 data.

9.3.3. Goodness-of-fit

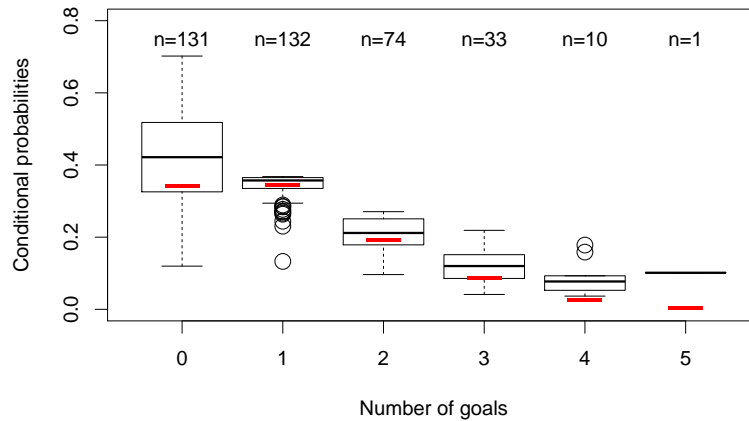
It is well-known that the scores of both competing teams in a soccer match are correlated. Several approaches to handle the correlation have been proposed in the literature. For example, in an unregularized setting McHale and Scarf (2006, 2011) model the dependence by using bivariate discrete distributions and by specifying a suitable family of dependence copulas. One of the first works investigating the topic of dependency between scores of competing soccer team is the fundamental article of Dixon and Coles (1997). They have shown that the joint distribution of the scores of both teams can not be well represented by the product of two independent marginal Poisson distributions of the home and away teams. They suggest to use an additional term to adjust for certain under- and overrepresented

match results. After all, these findings are based on the marginal distributions and only hold for models where the predictors of both scores are uncorrelated. However, the model proposed by Dixon and Coles (1997) includes team-specific attack and defense ability parameters and then uses independent poisson distributions for the numbers of goals scored. Therefore, the linear predictor for the number of goals of a specific team depends both on parameters of the team itself and its competitor. When fitting such a model to our World Cup data it turned out that the estimates of the attack and defense abilities of the teams are positively correlated. Therefore, although independent Poisson distributions are used for the scores in one match, the linear predictors and, accordingly, the predicted outcomes are (negatively) correlated. This holds both for the model of Dixon and Coles (1997) and, even more, for our proposed model where the linear predictors additionally entail covariates of both teams. To check if this phenomenon represents the actual correlations between the scores in one match in an appropriate manner, we compared the correlations between the real outcomes and the predictions from our model, exemplarily for the WC1994 data. We measured the correlation between 10,000 predictions for every match from the data set and compared it to the actual correlation between the scores in these matches. While we found a rank correlation (Spearman) of $\rho_{data} = -0.0882$ for the real outcomes, the predictions from our model have a rank correlation of $\rho_{model} = -0.0908$. The correlations according to Bravais-Pearson show similar results, $r_{data} = -0.1387$ and $r_{model} = -0.0968$. Alternatively, one can also investigate the residuals of the fitted model. If the model is representing the correlation structure in the data appropriately, the residuals belonging to the same match should be uncorrelated. For the WC1994 data we found correlations (accompanied by 95% bootstrap confidence intervals) according to Bravais-Pearson of 0.0198 (CI: [-0.0867;0.1283]) for the deviance residuals and of 0.0062 (CI: [-0.0977;0.1141]) for the Pearson residuals, respectively. In general, the point estimates show that the actual residuals of our model are uncorrelated. Still, due to the rather low number of observations, we obtain rather wide confidence intervals. Altogether, the correlations within the linear predictors for both teams competing in a match seem to fully account for the correlation between the scores of those teams and there is no need for further adjustment.

In a second step, we examined the actual distributions of the numbers of goals and compared them to the following (conditional) probabilities predicted by our model: separately for each plausible score from 0 to 5 goals we compared the observed proportion of the score in the data set with the probabilities for this score predicted by the model on only those observations showing this score. Figure 9.4 shows the corresponding boxplots, both using the WC1994 data (upper plot) and the WC2002 data including the odds (lower plot). The boxplots represent the probabilities of the respective scores predicted by our model, conditioned on those observations, whose actual number of goals equate to those scores. The red lines represent the relative frequencies of the respective scores in the data set. Note that if no statistical model is available the relative frequencies would serve as a natural, simple



(a) FIFA World Cup data 1994-2010

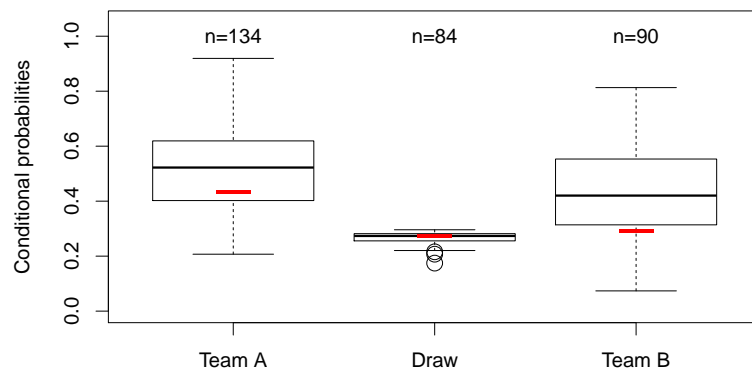


(b) FIFA World Cup data 2002-2010

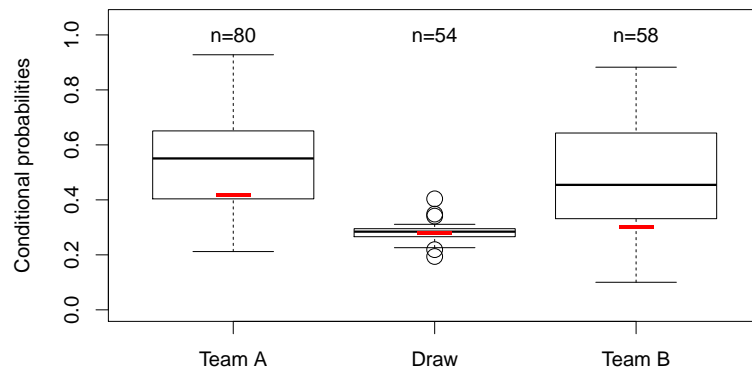
Figure 9.4.: Conditional probabilities of the numbers of goals predicted by the model for the FIFA World Cup data 1994-2010 (a) and by the model for the 2002-2010 data set (b). Red lines represent the relative frequencies of the respective scores in the data set and the corresponding absolute frequencies are displayed on top of every boxplot.

basis for the sampling of scores. So every statistical model should be able to compete with these relative frequencies in the sense that it should produce conditional predicted probabilities for each score exceeding these frequencies as far as possible. It can be seen that the model shows a good prediction performance regarding the number of goals. For example, for those 191 observations with an actual number of goals of zero, we observed a median of the conditional predicted probabilities of 36,1%, while the proportion in the data set for scores of zero was only 31,0%. In general, for all scores, the predicted conditional probabilities exceed the relative frequencies in the majority of cases. With respect to this criterion, the model for the data set including the odds (World Cups 2002-2010) performs slightly better than the model on the WC1994 data.

Another important aspect when modeling soccer matches based on (Poisson distributed) scores is a possible underestimation of draws, see e.g. Dixon and Coles (1997) and Karlis and Ntzoufras (2003). For the actual match outcome (i.e. win of team A , draw or win of team B) we performed an analysis similar to the different number of goals shown above. Separately for all three possible match outcomes we compared the relative frequencies of the outcome to the predicted probabilities of the respective true match outcome, conditioned on only those matches showing this outcome, see Figure 9.5. Interestingly, the first-mentioned



(a) FIFA World Cup data 1994-2010



(b) FIFA World Cup data 2002-2010

Figure 9.5.: Conditional probabilities of the actual match outcome (i.e. win team A , draw or win of team B) predicted by the model for the WC1994 data (a) and by the model for the WC2002 data (b). Red lines represent the relative frequencies of the respective outcomes in the data set and the corresponding absolute frequencies are displayed on top of every boxplot.

teams win more often than the second-mentioned teams. This is probably a consequence of the FIFA arrangement of the matches in the group stage and the round of sixteen. Hence, it seems reasonable to distinguish between wins of the “home” and “away” teams.

Although draws are generally predicted less well than wins of one of the teams, we found no systematic underestimation of draws. Again, the performance on the WC2002 data is slightly better.

9.3.4. Prediction Power

In the following, we try to assess the performance with respect to prediction of our model. At <http://www.oddsportal.com/soccer/world/world-cup-2014/results/> “three-way” odds⁶ for all 64 matches of the FIFA World Cup 2014, averaged over 16 well-known bookmakers, are provided. By taking the three quantities $\tilde{p}_r = 1/\text{odds}_r, r \in \{1, 2, 3\}$ and by normalizing with $c := \sum_{r=1}^3 \tilde{p}_r$ in order to adjust for the bookmakers’ margins, the odds can be directly transformed into probabilities using $\hat{p}_r = \tilde{p}_r/c$ ⁷. On the other hand, let G_{ij} denote the random variables representing the number of goals scored by team i in a certain match against team j and G_{ji} the goals of its opponent, respectively. Then, we can compute the same probabilities by approximating $\hat{p}_1 = P(G_{ij} > G_{ji}), \hat{p}_2 = P(G_{ij} = G_{ji})$ and $\hat{p}_3 = P(G_{ij} < G_{ji})$ for each of the 64 matches of the FIFA World Cup 2014 using the corresponding Poisson distributions $G_{ij} \sim \text{Poisson}(\hat{\lambda}_{ij}), G_{ji} \sim \text{Poisson}(\hat{\lambda}_{ji})$, where the estimates $\hat{\lambda}_{ij}$ and $\hat{\lambda}_{ji}$ are obtained by our regression models. Based on these predicted probabilities, the average probability of a correct prediction of a FIFA World Cup 2014 match can be obtained. For the true match outcomes $\omega_m \in \{1, 2, 3\}, m = 1, \dots, 64$, it is given by $\bar{p}_{\text{three-way}} := \frac{1}{64} \sum_{m=1}^{64} \hat{p}_{1m}^{\delta_{1\omega_m}} \hat{p}_{2m}^{\delta_{2\omega_m}} \hat{p}_{3m}^{\delta_{3\omega_m}}$, with δ_{rm} denoting Kronecker’s delta. The quantity $\bar{p}_{\text{three-way}}$ serves as a useful performance measure for a comparison of the predictive power of the model and the bookmakers’ odds and is shown for both data sets in Table 9.4. It is striking that the predictive power of our model compares well with the bookmakers’ odds for both data sets, especially if one has in mind that the bookmakers odds are usually released just some days before the corresponding match takes place and, hence, are able to include the latest performance trends of both competing teams. In general, the out-of-sample prediction seems very satisfying to us, with slightly better results for the WC2002 data.

If one puts one’s trust into the model and its predicted probabilities, it could serve as the basis of the following betting strategy: for every match one would bet on the three-way match outcome with the highest expected return, which can be calculated as the product of the model’s predicted probability and the corresponding three-way odd offered by the bookmakers. We applied this strategy to the model results of both data sets, yielding a

⁶ Three-way odds consider only the tendency of a match with the possible results *victory of team 1, draw or defeat of team 1* and are usually fixed some days before the corresponding match takes place.

⁷ The transformed probabilities only serve as an approximation, based on the assumption that the bookmakers’ margins follow a discrete uniform distribution on the three possible match tendencies.

return of 33.52% for WC2002 and 19.28% for WC1994, when for all 64 matches equal-sized bets are placed. Again, this is a very satisfying result with an advantage for WC2002.

WC1994	WC2002	bookmakers' odds
40.15%	40.33%	41.45%

Table 9.4.: Average probability $\bar{p}_{\text{three-way}}$ of a correct prediction of a FIFA World Cup 2014 match for our model on both data sets and the bookmakers' odds.

In Table 9.5, the corresponding estimates of the (unscaled) fixed team-specific attacking and defending effects are summarized, exemplarily for the WC2002 data. In contrast to the covariate effects from Table 9.3, we present the unscaled effects here, as this allows a direct comparison of both the attack and defense parameters of different teams. As already pointed out in Section 9.2, the full attack or defense abilities of team i are represented by the terms $\mathbf{x}_{ik}^T \boldsymbol{\beta} + att_i$ and $\mathbf{x}_{ik}^T \boldsymbol{\beta} + def_i$, respectively, and not only by the parameters att_i and def_i . Therefore, $att_i = def_i = 0$ simply indicates that for such teams no additional attack or defense effects are needed. In general, larger team-specific attack or defense parameters, respectively, increase the team's performance. It is striking that compared to all other teams Germany and Brazil both have rather high attacking and defending abilities: Germany's attack is on 1st place, its defense is on 3rd place; Brazil's attack is on 2nd place, its defense on 5th place. In this context, also the parameters of Switzerland are interesting. Switzerland has a rather bad attack, but the best defense parameter among all the teams. This can be easily explained, as Switzerland has received only a single goal in its seven games at the World Cups 2006 and 2010, but on the other hand only scored five goals in these seven matches. Table 9.5 also provides the exponentials of the ability parameters. Due to the used (log-)link, they represent the multiplicative (or divisive) effects of the respective parameters on the response scale. In the current example, this means that the number of goals Switzerland concedes are divided by 1.8 compared to the case where Switzerland would not have an additional defense parameter.

9.3.5. Probabilities for FIFA World Cup 2014 Winner

We used both estimates from the two models fitted on the WC1994 and the WC2002 data to simulate the tournament progress of the FIFA World Cup 100,000 times. As we have seen above that the model on the WC2002 data performs slightly better than the WC1994 model with respect to all regarded goodness-of-fit and prediction criteria, we present in this section only the prediction results of the model based on the WC2002 data. The results corresponding to the WC1994 data can be found in Appendix C.















estimated attack parameters					estimated defense parameters				
1.		GER	0.237	1.267	1.		SUI	0.599	1.821
2.		BRA	0.114	1.121	2.		ALG	0.205	1.227
3.		URU	0.101	1.106	3.		GER	0.181	1.199
4.		CRC	0.099	1.104	4.		HON	0.065	1.067
5.		RSA	0.060	1.062	5.		BRA	0.057	1.059
6.		BEL	0.042	1.043	6.		FRA	0.046	1.047
7.		POR	0.019	1.019	7.		POR	0.030	1.031
8.		ARG	0.000	1.000	8.		PAR	0.021	1.021
9.		AUS	0.000	1.000	9.		ARG	0.000	1.000
10.		CHI	0.000	1.000	10.		AUS	0.000	1.000
11.		CRO	0.000	1.000	11.		CHI	0.000	1.000
12.		DEN	0.000	1.000	12.		CRO	0.000	1.000
13.		ECU	0.000	1.000	13.		DEN	0.000	1.000
14.		ENG	0.000	1.000	14.		ECU	0.000	1.000
15.		GHA	0.000	1.000	15.		ENG	0.000	1.000
16.		GRE	0.000	1.000	16.		GHA	0.000	1.000
17.		ITA	0.000	1.000	17.		GRE	0.000	1.000
18.		CIV	0.000	1.000	18.		ITA	0.000	1.000
19.		JPN	0.000	1.000	19.		CIV	0.000	1.000
20.		MEX	0.000	1.000	20.		JPN	0.000	1.000
21.		NED	0.000	1.000	21.		MEX	0.000	1.000
22.		NEW	0.000	1.000	22.		NED	0.000	1.000
23.		NGA	0.000	1.000	23.		NEW	0.000	1.000
24.		RUS	0.000	1.000	24.		NGA	0.000	1.000
25.		KOR	0.000	1.000	25.		RUS	0.000	1.000
26.		ESP	0.000	1.000	26.		KOR	0.000	1.000
27.		SWE	0.000	1.000	27.		ESP	0.000	1.000
28.		USA	0.000	1.000	28.		SWE	0.000	1.000
29.		SVN	-0.002	0.998	29.		USA	0.000	1.000
30.		PAR	-0.003	0.997	30.		SVN	-0.009	0.991
31.		POL	-0.005	0.995	31.		POL	-0.012	0.988
32.		IRN	-0.040	0.960	32.		URU	-0.019	0.981
33.		SRB	-0.047	0.954	33.		RSA	-0.022	0.978
34.		HON	-0.083	0.921	34.		BEL	-0.085	0.918
35.		FRA	-0.198	0.821	35.		CMR	-0.090	0.914
36.		SUI	-0.202	0.817	36.		IRN	-0.097	0.907
37.		CMR	-0.204	0.815	37.		SRB	-0.153	0.858
38.		TUN	-0.234	0.791	38.		TUN	-0.297	0.743
39.		ALG	-0.340	0.712	39.		CRC	-0.526	0.591
40.		KSA	-0.495	0.610	40.		KSA	-0.788	0.455

Table 9.5.: (Unscaled) estimates of the team-specific attacking effects att_i and their exponentials $\exp(att_i)$ (left) and defending effects def_i and their exponentials $\exp(def_i)$ (right) for the WC2002 data.

Note here that one advantage in comparison to several other prediction approaches is that we are able to draw exact match outcomes for each match by drawing the goals of both competing teams from the predicted Poisson distributions, i.e. $G_{ij} \sim \text{Poisson}(\hat{\lambda}_{ij})$, $G_{ji} \sim \text{Poisson}(\hat{\lambda}_{ji})$, with estimates $\hat{\lambda}_{ij}$ and $\hat{\lambda}_{ji}$ from the WC2002 model. This allows us to precisely follow the official FIFA rules when determining the final group standings⁸. If a match in the knockout stage ended in a draw, we simulated another 30 minutes of extra time using scoring rates equal to 1/3 of the 90 minutes rates. If the match then still ended in a draw, the winner was calculated simply by coin flip, reflecting a penalty shoot out.

Based on these simulations, for each of the 32 participating teams probabilities to reach the next stage and, finally, to win the tournament are obtained. These are summarized in Table 9.6 together with the winning probabilities based on the ODDSET odds for comparison. In contrast to most other prediction approaches for the FIFA World Cup 2014 clearly favoring Brazil, we get a neck-and-neck race between Germany and Brazil, finally with better chances for Germany. The major reason for this is that with a high probability in the simulations both Germany and Brazil finish their groups on the first place and then face each other in the semi final. In a direct duel, the model concedes Germany a thin advantage with a winning probability of 51,7% against 48,3%. The favorites Germany and Brazil are followed by the teams of Switzerland, Spain, Argentina and Portugal. Similarly, for the WC1994 data Germany has the highest probability to win the trophy, followed by Spain and Brazil, see Table C.1 in Appendix C.

In a second step, we investigate how the model (and the respective winning probabilities) change when the data set is successively extended by the completed matches of the current World Cup in each stage. For example, after the group stage the model is refitted including all 48 matches from the group stage. Then, for the round of 16 the qualified teams from the group stage are known and used for the prediction of the round of 16. For example, according to the initial model Costa Rica appeared to be a clear underdog and only had low chances to reach the round of 16 (7.7%). Based on the initial model, in the upcoming knockout match against Greece, Costa Rica's probability to qualify for the quarter finals was estimated to be 27.8%, whereas the adapted model yields an increased probability of 42.8%. Therefore, the model accounted for the good performance of Costa Rica in the group stage and, indeed, Costa Rica actually defeated Greece in a penalty shootout. A similar effect comes up for the following quarter final between Costa Rica and the Netherlands where the chances of Costa Rica are increased from 19.3% to 32.9%. Again, the real match was actually quite close with Netherlands winning in another penalty shootout. Table 9.7 summarises the adapted probabilities for all stages, again based on 100,000 simulation runs.

⁸ The final group standings are determined by (1) the number of points, (2) the goal difference and (3) the number of scored goals. If several teams coincide with respect to all of these three criteria, a separate chart is calculated based on the matches between the coinciding teams only. Here, again the final standing of the teams is determined following criteria (1)-(3). If still no distinct decision can be taken, the decision is taken by lot.

































			Round of 16	Quarter finals	Semi finals	Final	World Champion	Oddset
1.		GER	91.4	77.9	57.0	39.2	27.6	14.2
2.		BRA	91.8	67.9	54.4	30.9	20.0	20.3
3.		SUI	84.2	62.0	35.0	21.6	12.5	0.7
4.		ESP	84.2	52.0	37.8	21.6	12.1	10.9
5.		ARG	90.6	53.2	26.7	15.5	7.3	14.2
6.		POR	60.2	38.6	20.2	10.4	3.6	2.4
7.		BEL	82.5	36.3	19.8	9.3	3.4	5.9
8.		ENG	70.4	41.2	14.7	5.5	1.8	3.5
9.		CRO	58.1	26.1	15.5	5.1	1.6	0.7
10.		FRA	51.2	26.5	9.8	4.6	1.1	3.5
11.		ITA	56.8	31.8	10.8	4.4	1.3	3.5
12.		NED	55.7	21.3	11.8	4.1	1.2	3.5
13.		URU	65.1	37.3	11.6	4.1	1.2	2.8
14.		COL	60.6	31.5	10.7	4.0	1.2	3.9
15.		CIV	58.3	26.3	9.2	2.9	0.6	0.7
16.		CHI	42.9	13.5	7.3	2.5	1.0	2.0
17.		GRE	53.2	22.5	7.3	2.1	0.4	0.7
18.		USA	27.2	13.0	5.4	2.0	0.4	0.7
19.		MEX	42.0	14.7	5.8	1.9	0.3	0.7
20.		GHA	21.2	8.7	4.0	1.8	0.5	0.7
21.		RUS	51.3	11.5	4.3	1.4	0.3	1.2
22.		HON	28.3	10.5	3.6	1.2	0.3	0.1
23.		KOR	41.4	9.7	3.4	1.0	0.0	0.2
24.		BIH	48.2	17.1	3.8	0.8	0.1	0.5
25.		ECU	36.3	14.8	3.4	0.8	0.1	0.7
26.		JPN	27.9	7.7	1.7	0.4	0.0	0.5
27.		ALG	24.8	4.3	1.1	0.4	0.1	0.1
28.		NGA	39.4	12.5	2.1	0.2	0.0	0.4
29.		IRN	21.8	3.4	0.4	0.2	0.0	0.1
30.		AUS	17.2	2.8	0.8	0.1	0.0	0.2
31.		CMR	8.1	1.7	0.5	0.0	0.0	0.2
32.		CRC	7.7	1.7	0.1	0.0	0.0	0.1

Table 9.6.: Estimated probabilities (in %) for reaching the different stages in the FIFA World Cup 2014 for all 32 teams based on 100,000 simulation runs of the FIFA World Cup 2014 and based on the estimates of the WC2002 data together with winning probabilities based on the ODDSET odds.

































			Round of 16	Quarter finals	Semi finals	Final	World Champion
1.		GER	91.4	78.5	71.2	48.7	72.5
2.		ARG	90.6	58.9	49.9	57.6	27.5
3.		BRA	91.8	76.7	54.2	51.3	0.0
4.		NED	55.7	59.2	67.1	42.4	0.0
5.		BEL	82.5	65.6	50.1	0.0	0.0
6.		COL	60.6	81.1	45.8	0.0	0.0
7.		CRC	7.7	42.8	32.9	0.0	0.0
8.		FRA	51.2	71.8	28.8	0.0	0.0
9.		GRE	53.2	57.2	0.0	0.0	0.0
10.		SUI	84.2	41.1	0.0	0.0	0.0
11.		MEX	42.0	40.8	0.0	0.0	0.0
12.		USA	27.2	34.4	0.0	0.0	0.0
13.		NGA	39.4	28.2	0.0	0.0	0.0
14.		CHI	42.9	23.3	0.0	0.0	0.0
15.		ALG	24.8	21.5	0.0	0.0	0.0
16.		URU	65.1	18.9	0.0	0.0	0.0
17.		ESP	84.2	0.0	0.0	0.0	0.0
18.		ENG	70.4	0.0	0.0	0.0	0.0
19.		POR	60.2	0.0	0.0	0.0	0.0
20.		CIV	58.3	0.0	0.0	0.0	0.0
21.		CRO	58.1	0.0	0.0	0.0	0.0
22.		ITA	56.8	0.0	0.0	0.0	0.0
23.		RUS	51.3	0.0	0.0	0.0	0.0
24.		BIH	48.2	0.0	0.0	0.0	0.0
25.		KOR	41.4	0.0	0.0	0.0	0.0
26.		ECU	36.3	0.0	0.0	0.0	0.0
27.		HON	28.3	0.0	0.0	0.0	0.0
28.		JPN	27.9	0.0	0.0	0.0	0.0
29.		IRN	21.8	0.0	0.0	0.0	0.0
30.		GHA	21.2	0.0	0.0	0.0	0.0
31.		AUS	17.2	0.0	0.0	0.0	0.0
32.		CMR	8.1	0.0	0.0	0.0	0.0

Table 9.7.: Estimated (adapted) probabilities (in %) for reaching the next stages in the FIFA World Cup 2014 for all 32 teams based on 100,000 simulation runs of the FIFA World Cup 2014. After each round, the data set (WC2002) is extended with by the matches already played and the model is refitted. Only actual matches from the World Cup are simulated.

In Appendix C, Table C.2 shows the respective (adapted) probabilities for the WC1994 data.

9.3.6. Most Probable Tournament Outcome

Finally, based on the 100,000 simulations, we also provide the most probable tournament outcome, exemplarily for the WC2002 data. Here, for each of the eight groups we selected the most probable final group standing, also regarding the order of the first two places, but without regarding the irrelevant order of the teams on place three and four. The results together with the corresponding probabilities are presented in Table 9.8.

Group A 39%	Group B 26%	Group C 15%	Group D 19%
1.  BRA	1.  ESP	1.  COL	1.  ENG
2.  CRO	2.  NED	2.  GRE	2.  ITA
 MEX	 CHI	 JPN	 URU
 CMR	 AUS	 CIV	 CRC
Group E 19%	Group F 29%	Group G 38%	Group H 23%
1.  SUI	1.  ARG	1.  GER	1.  BEL
2.  FRA	2.  BIH	2.  POR	2.  RUS
 ECU	 NGA	 GHA	 ALG
 HON	 IRN	 USA	 KOR

Table 9.8.: Most probable final group standings together with the corresponding probabilities for the FIFA World Cup 2014 based on 100,000 simulation runs and on the estimates of the WC2002 data.

It is obvious that there are large differences with respect to the groups' balances. While in Group A and Group G the model forecasts Brazil followed by Croatia as well as Germany followed by Portugal with rather high probabilities of 39% and 38%, respectively, other groups such as Group C, Group D and Group E seem to be quite close.

Based on the most probable group standings, we also provide the most probable course of the knockout stage, compare Figure 9.6. Finally, according to the most probable tournament course the German team will take home the World Cup trophy. Although according to the model this reflects the most probable tournament outcome, it only has a very low overall probability of $1.49 \cdot 10^{-6} \%$ (given as the product of all single probabilities of Table 9.8 and Figure 9.6). Hence, deviations of the true tournament outcome from the model's most probable one are not only possible, but very likely.

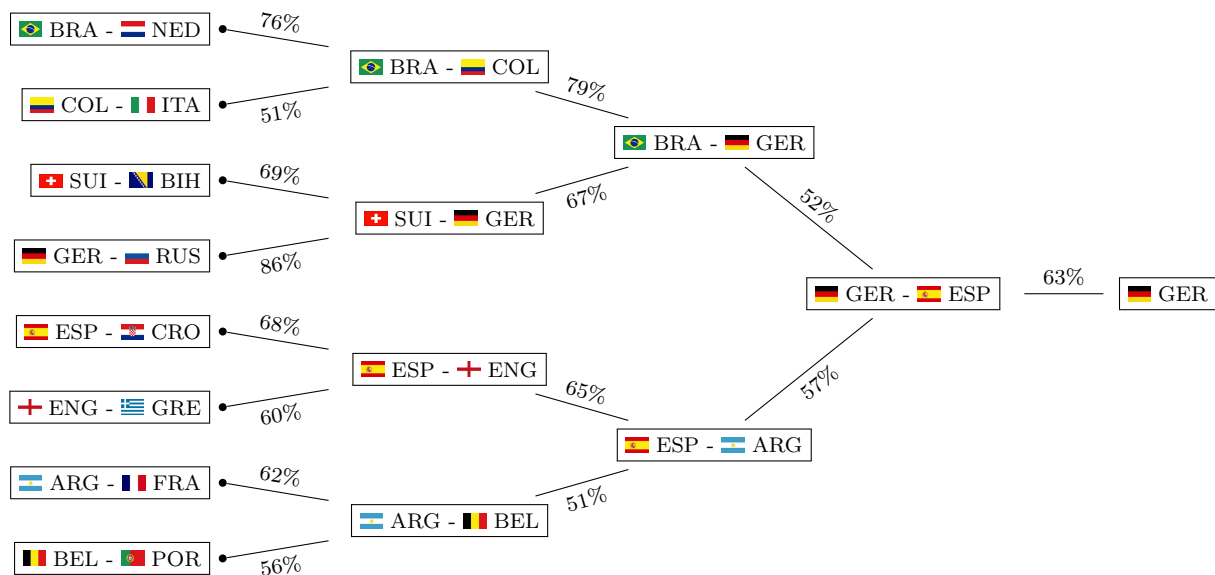


Figure 9.6.: Most probable course of the knockout stage together with corresponding probabilities for the FIFA World Cup 2014 based on 100,000 simulation runs and on the estimates of the WC2002 data.

In fact, if we compare the most probable tournament outcome of the FIFA World Cup 2014 from Table 9.8 and Figure 9.6 with the true one, several differences become obvious. In general, several underdogs, such as e.g. Algeria, Costa Rica, USA or Chile have reached the round of sixteen, while several favorites, such as e.g. Spain, Italy, England or Portugal, dropped out already in the group stage. This could not be adequately represented by the model. Nevertheless, beyond the round of sixteen, the model's predicted tournament course gets closer and closer to the true one, with three out of four semi-finalists predicted correctly and finally, with Germany correctly predicted as the World Champion.

9.4. Concluding Remarks

A team-specific generalized linear Poisson model for the number of goals scored by soccer teams facing each other in international tournament matches is set up. As an application, the FIFA World Cups 1994-2010 and 2002-2010, respectively, serve as the data basis for an analysis of the influence of several covariates on the success of national teams in terms of the number of goals they score in single matches. Procedures for variable selection based on an L_1 -penalty, implemented in the R-package `grplasso`, are used. A detailed goodness-of-fit analysis is presented and suitable “out-of-sample” performance measures for prediction are considered, which are based on the three-way tendencies of the considered matches.

The fitted models were used for simulation of the FIFA World Cup 2014. According to these simulations, Germany and Brazil turned out to be the top favorites for winning the title, with an advantage for Germany. Besides, the most probable tournament outcome is provided.

A major part of the statistical novelty of the presented work lies in the use of penalty terms for covariate effects in combination with team-specific abilities. It allows to include many covariates simultaneously and performs automatic variable selection. In the case of high correlation between certain covariates, the estimation procedure is stabilized by the penalization. If several high correlated variables possibly contain information on the response, the LASSO tends to include the predictor with the highest explanatory power. Furthermore, as the basic model used throughout this chapter is in general not identified, the penalized likelihood approach nevertheless allows for unique estimates. Theoretically, this would also allow for the estimation of effects of covariates not varying over different tournaments, which are un-separable from team-specific effects in an unpenalized estimation.

Another important aspect is that the team-specific ability parameters need not necessarily be constant, but instead could evolve over time since composition and performance of the teams might change over time. In this context we want to mention a very recent publication of Koopman and Lit (2015). They assume a bivariate Poisson distribution for the goals in English Premier League matches, with intensity coefficients that change stochastically over time by modeling the teams’ ability parameters as first order auto-regressive processes. However, due to certain general differences in the structure of national league and FIFA World Cup data it is not straightforward, how this approach can be adopted to the present data situation. Nevertheless, the idea of time-varying ability parameters in modeling international soccer data sounds promising to us and could be the starting point for future research.

10. Conclusion and Outlook

The focus of this thesis is on the extension of models for heterogeneous (i.e. item response data) and homogeneous paired comparisons, mainly by the inclusion of different kinds of covariates. In particular, the Rasch model and the Bradley-Terry model are used as basis and extended in various ways.

Extensions of IRT Models

In the first part of this concluding chapter, the basic Rasch model and the proposed extension of the Rasch model are recapitulated. Furthermore, possible further extensions of different IRT models are considered. Following the notation from Chapter 2, the Rasch model is denoted by

$$P(Y_{pi} = 1) = \frac{\exp(\theta_p - \beta_i)}{1 + \exp(\theta_p - \beta_i)}$$

when person p and item i is considered. In Chapters 4 and 5, the Rasch model is extended by a term considering (subject-specific) covariates \mathbf{x}_p . The resulting model, referred to as the DIF model, is denoted by

$$P(Y_{pi} = 1|\mathbf{x}_p) = \frac{\exp(\theta_p - (\beta_i + \mathbf{x}_p^T \boldsymbol{\gamma}_i))}{1 + \exp(\theta_p - (\beta_i + \mathbf{x}_p^T \boldsymbol{\gamma}_i))}. \quad (10.1)$$

The model is called DIF model as it can be used to detect items with differential item functioning. For that purpose, two different estimation strategies were proposed. Both strategies are based on regularization techniques, namely penalization and boosting. The regularization techniques allow for feature selection and, therefore, are able to select the DIF items.

Within the item response theory, the Rasch model is a very popular yet quite restrictive model. It can be seen as special case of the general Birnbaum model or 3PL model which is denoted by

$$P(Y_{pi} = 1) = c_i + (1 - c_i) \frac{\exp(a_i(\theta_p - \beta_i))}{1 + \exp(a_i(\theta_p - \beta_i))}.$$

Instead of only one parameter characterizing an item, the 3PL model has 3 item parameters. Beside the item difficulty β_i , it contains the discrimination parameter a_i and the guessing parameter c_i . In the Rasch model (or 1PL model), the restrictions $a_i = 1$ and $c_i = 0$ are applied. In the so-called 2PL model, the discrimination parameters a_i can vary and only $c_i = 0$ is applied. An obvious extension of the model proposed in Chapters 4 and 5 would be to use the 2PL or even the 3PL and to extend it by covariate effects.

In particular, extending the 2PL model could be an interesting tool for DIF analysis. In the proposed DIF model, the item difficulty β_i is replaced by the term $\beta_i + \mathbf{x}_p^T \boldsymbol{\gamma}_i$ to identify uniform DIF. If (in the 2PL model) the discrimination parameter a_i would be replaced by the term $a_i + \mathbf{x}_p^T \boldsymbol{\delta}_i$, such a model could be used to identify non-uniform DIF. Non-uniform is characterized by item characteristic curves with different slopes, for example between two subgroups of the population like males and females. Therefore, if x_p would simply refer to gender, any estimate $\delta_i \neq 0$ would indicate that item i has non-uniform DIF with respect to the subgroups males and females. Consequently, similar to the concept proposed in Chapters 4 and 5 this idea could be extended to a whole vector of person-specific covariates \mathbf{x}_p possibly containing both categorical and continuous covariates. Clearly, such a model would be a challenge in terms of estimation. Because of its multiplicative form, the 2PL model can not be embedded into the framework of generalized linear models. Typically, a marginal likelihood approach is chosen, see also Section 2.2. This estimation concept could be combined with a penalty approach similar to Chapter 4 for an automatic selection of items with non-uniform DIF. Furthermore, also boosting concepts similar to Chapter 5 could be applied possibly circumventing the problem of the multiplicative nature of the linear predictor.

Another possible extension of the proposed DIF model (10.1) could be to include item-specific covariates. Let us consider the knowledge data from Subsection 4.5.2. For the DIFlasso method, person-specific covariates were used to find DIF items with respect to these covariates. The items were supposed to measure the latent trait general knowledge. The items can be divided into five topics, namely politics, history, economy, culture, and natural sciences. Therefore, the items themselves have special characteristics that could influence the probability that a person solves a specific item. A possible extension of model (10.1) by item-specific covariates could be denoted by

$$P(Y_{pi} = 1 | \mathbf{x}_p, \mathbf{z}_i) = \frac{\exp(\theta_p - (\beta_i + \mathbf{x}_p^T \boldsymbol{\gamma}_i + \mathbf{z}_i^T \boldsymbol{\delta}))}{1 + \exp(\theta_p - (\beta_i + \mathbf{x}_p^T \boldsymbol{\gamma}_i + \mathbf{z}_i^T \boldsymbol{\delta}))}.$$

If \mathbf{z}_i contains the information on the topic of item i (in dummy coding), $\boldsymbol{\delta}$ would contain parameters characterizing a general level of difficulty of the single topics. The model is not identifiable as in principle the terms β_i and $\mathbf{z}_i^T \boldsymbol{\delta}$ are not separable. Therefore, for such a model additional restrictions or an additional penalty term have to be applied.

Theoretically, the global effects δ could be replaced by person-specific effects resulting in the model

$$P(Y_{pi} = 1 | \mathbf{x}_p, \mathbf{z}_i) = \frac{\exp(\theta_p - (\beta_i + \mathbf{x}_p^T \gamma_i + \mathbf{z}_i^T \delta_p))}{1 + \exp(\theta_p - (\beta_i + \mathbf{x}_p^T \gamma_i + \mathbf{z}_i^T \delta_p))}.$$

Applied to the topics of the items, in this model every person would have its own difficulty parameter for every topic. Therefore, one could for example see if a specific person performs better on items in history or natural science. However, this would involve a tremendous increase of the number of parameters, namely one parameter per person and item-specific covariate. Therefore, even if appropriate penalty terms were applied such a model would probably lead to problems concerning computation time and especially interpretability.

Extensions of Paired Comparison Models

In the second part of this chapter, the proposed extensions of paired comparison models and possible future extensions are discussed. In the following, the case of ordered paired comparisons is skipped for the sake of simplicity. The respective extensions are straightforward. According to the notation from Chapter 6, the basic Bradley-Terry model can be denoted by

$$P(Y_{(r,s)} = 1) = \frac{\exp(\gamma_r - \gamma_s)}{1 + \exp(\gamma_r - \gamma_s)}$$

considering a comparison between objects a_r and a_s .

Inclusion of Subject-specific Covariates

In this thesis, the Bradley-Terry model is extended in two different ways. First, in Chapter 7 the Bradley-Terry model is extended by the inclusion of subject-specific covariates. The respective application considered data on party preference from Germany. As the data originate from the party preference of different persons, obviously subject-specific covariates are attributes of the interviewed persons like age, gender or educational level. Therefore, the extended model can be denoted by

$$\begin{aligned} P(Y_{i(r,s)} = 1 | \mathbf{x}_i) &= \frac{\exp(\gamma_{ir} - \gamma_{is})}{1 + \exp(\gamma_{ir} - \gamma_{is})} \\ \text{with } \gamma_{ir} &= \beta_{r0} + \mathbf{x}_i^T \boldsymbol{\beta}_r \end{aligned} \quad (10.2)$$

where \mathbf{x}_i contains characteristics of subject i . In this model, the subject-specific covariates are connected to object-specific parameters $\boldsymbol{\beta}_r = (\beta_{r1}, \dots, \beta_{rp})$. To control for the increasing complexity of the model, penalty terms are used for the estimation. The object-specific

parameters are penalized using the penalty term $J(\boldsymbol{\alpha}) = \sum_j \sum_{r < s} w_{rsj} |\beta_{rj} - \beta_{sj}|$. The respective parameters are shrunk toward each other and can be merged or even be eliminated from the model.

Inclusion of Object-specific Covariates

Second, in Chapter 8 the Bradley-Terry model is extended by the inclusion of object-specific covariates. The respective model can be denoted by

$$\begin{aligned} P(Y_{(r,s)} = 1 \mid \mathbf{z}_r, \mathbf{z}_s) &= \frac{\exp(\gamma_r - \gamma_s)}{1 + \exp(\gamma_r - \gamma_s)} \\ \text{with } \gamma_r &= \beta_{r0} + \mathbf{z}_r^T \boldsymbol{\alpha} \end{aligned} \quad (10.3)$$

where \mathbf{z}_r contains characteristics of object a_r . Exemplarily, this extension is done for a model for the German Bundesliga. The competing objects in this model are the respective football clubs. The object-specific covariate we consider is the budget of the clubs. The main difference between the inclusion of subject-specific and object-specific covariates is that the effects of object-specific covariates are not separable from the regular strength/attractivity parameters β_{r0} of the objects. Therefore, such a model is not identifiable. The model in Chapter 8 is made estimable by the penalty term $J(\boldsymbol{\beta}) = \sum_{r < s} w_{rs} |\beta_{r0} - \beta_{s0}|$. Beside providing a unique solution of the estimation problem, the penalty term shrinks the strength parameters of the teams toward each other and can find clusters of teams with the same strength parameters.

Combining Subject-specific and Object-specific Covariates

An obvious extension of the proposed models would be to include subject-specific and object-specific covariates into a model simultaneously. Clearly, for such a model the estimation techniques, or rather the penalty terms, will depend on the respective application. Exemplarily, these differences shall be discussed in the following with respect to the applications from Chapters 7 and 8.

Combining models (10.3) and (10.4), a general model including both subject-specific and object-specific covariates could be denoted by

$$\begin{aligned} P(Y_{i(r,s)} = 1 \mid \mathbf{x}_i, \mathbf{z}_r, \mathbf{z}_s) &= \frac{\exp(\gamma_{ir} - \gamma_{is})}{1 + \exp(\gamma_{ir} - \gamma_{is})} \\ &= \frac{\exp(\beta_{r0} - \beta_{s0} + \mathbf{x}_i^T (\boldsymbol{\beta}_r - \boldsymbol{\beta}_s) + (\mathbf{z}_r - \mathbf{z}_s)^T \boldsymbol{\alpha})}{1 + \exp(\beta_{r0} - \beta_{s0} + \mathbf{x}_i^T (\boldsymbol{\beta}_r - \boldsymbol{\beta}_s) + (\mathbf{z}_r - \mathbf{z}_s)^T \boldsymbol{\alpha})} \\ \text{with } \gamma_{ir} &= \beta_{r0} + \mathbf{x}_i^T \boldsymbol{\beta}_r + \mathbf{z}_r^T \boldsymbol{\alpha}. \end{aligned} \quad (10.4)$$

Beside the fact that in Chapter 7 subject-specific covariates are used while in Chapter 8 object-specific covariates are used, another fundamental difference exists, namely the number of objects. While in the party preference data only five different objects appear, for the German Bundesliga 18 objects have to be distinguished. Therefore, for the party preference data it is neither necessary nor desirable to reduce the complexity of the main effects β_{r0} in model (10.5) by finding clusters within the objects. In contrast, this is a sensible strategy for the application on the Bundesliga data.

Both for the party preference data and for the Bundesliga data a combination of subject-specific and objects-specific covariates is conceivable. For the party preference data, object-specific covariates would have to be covariates characterizing the respective parties. For example, one could include the number of party members, a variable indicating if the party currently is a governing party or a opposition party or a variable indicating if the leading candidate of the party is male or female. As long as the number of object-specific covariates is rather small, the respective parameters probably would not need to be penalized.

In the case of the Bundesliga data, subject-specific covariates would be covariates characterizing the respective match or match-day. Exemplarily, the weather conditions, the weekday or the number of spectators could be considered. Every considered covariate would have a separate parameter per team. Therefore, a penalty term equal to the penalty proposed for the subject-specific covariates in the party preference data seems mandatory.

Inclusion of Subject-object-specific Covariates

In the case of the party preference data, beside subject-specific and object-specific covariates even a third kind of covariates is conceivable, possibly called subject-object-specific covariates. Subject-object-specific covariates differ both between subjects and between objects. In the original pre-election data set GLES (Rattinger et al., 2014) used in Chapter 7 such variables appear. For example, the participants are asked about certain political topics like the topic of climate change. The respondents are supposed to report both their self-perception and their perception of the single parties toward this topic on a Likert scale with 11 ordered levels, corresponding to the following statements:

Level 1: *Fight against climate change should take precedence, even if it impairs economic growth.*

Level 11: *Economic growth should take precedence, even if it impairs the fight against climate change.*

The absolute difference between these perceptions can be seen as the (self-perceived) absolute distance between the person (subject) and the parties (objects) with respect to a

political topic. A model combining the subject-specific covariates from Chapter 7 with such subject-object-specific covariates can be denoted by

$$\begin{aligned}
 P(Y_{i(r,s)} = 1 \mid \mathbf{x}_i, \mathbf{z}_{ir}, \mathbf{z}_{is}) &= \frac{\exp(\gamma_{ir} - \gamma_{is})}{1 + \exp(\gamma_{ir} - \gamma_{is})} \\
 &= \frac{\exp(\beta_{r0} - \beta_{s0} + \mathbf{x}_i^T(\boldsymbol{\beta}_r - \boldsymbol{\beta}_s) + (\mathbf{z}_{ir} - \mathbf{z}_{is})^T \boldsymbol{\alpha})}{1 + \exp(\beta_{r0} - \beta_{s0} + \mathbf{x}_i^T(\boldsymbol{\beta}_r - \boldsymbol{\beta}_s) + (\mathbf{z}_{ir} - \mathbf{z}_{is})^T \boldsymbol{\alpha})} \\
 \text{with } \gamma_{ir} &= \beta_{r0} + \mathbf{x}_i^T \boldsymbol{\beta}_r + \mathbf{z}_{ir}^T \boldsymbol{\alpha}.
 \end{aligned} \tag{10.5}$$

Similar to object-specific covariates, only one parameter per covariate is necessary. Therefore, for a rather small number of covariates the inclusion of a further penalty term is possible but not mandatory. For the exemplary covariate climate change, the respective parameter could be interpreted as relevance of the topic climate change. The term $\mathbf{z}_{ir} - \mathbf{z}_{is}$ contains the difference between the absolute distances of the parties a_r and a_s and the subject i toward climate change. If this difference is positive, party a_r has a higher distance to person i than party a_s . Therefore, the respective effect α will probably be negative indicating that a person will rather prefer a party with a position close to the self-perception of the person, and vice versa. Parameters close to zero indicate that the topic is not very relevant for the decision between parties while extremely negative parameters indicate a high relevance of the topic.

Instead of including one global parameter for all parties, such an effect could also be party-specific. Then, party-specific effects are estimated representing the effect of the distance between a person and a party. Now, the effect of the position toward climate change is not equal but may vary between parties possibly showing that the topic climate change has a different relevance for different parties. Such a model can be seen as a further extension of model (10.5) which can be denoted by

$$\begin{aligned}
 P(Y_{i(r,s)} = 1 \mid \mathbf{x}_i, \mathbf{z}_{ir}, \mathbf{z}_{is}) &= \frac{\exp(\gamma_{ir} - \gamma_{is})}{1 + \exp(\gamma_{ir} - \gamma_{is})} \\
 &= \frac{\exp(\beta_{r0} - \beta_{s0} + \mathbf{x}_i^T(\boldsymbol{\beta}_r - \boldsymbol{\beta}_s) + \mathbf{z}_{ir}^T \boldsymbol{\alpha}_r - \mathbf{z}_{is}^T \boldsymbol{\alpha}_s)}{1 + \exp(\beta_{r0} - \beta_{s0} + \mathbf{x}_i^T(\boldsymbol{\beta}_r - \boldsymbol{\beta}_s) + \mathbf{z}_{ir}^T \boldsymbol{\alpha}_r - \mathbf{z}_{is}^T \boldsymbol{\alpha}_s)} \\
 \text{with } \gamma_{ir} &= \beta_{r0} + \mathbf{x}_i^T \boldsymbol{\beta}_r + \mathbf{z}_{ir}^T \boldsymbol{\alpha}_r.
 \end{aligned} \tag{10.6}$$

Here, subject-object-specific covariates are considered together with object-specific parameters. Implicitly, subject-object-specific covariates have already been considered in this thesis, both with global and object-specific parameters. In Chapter 8, different order effects were incorporated, in case of the Bundesliga application also referred to as home advantages. In particular, a global home advantage and team-specific home advantages were considered. A home advantage can be encoded by the subject-object-specific covariate $\mathbf{z}_{ir} = 1$ for the home team a_r in match i and $\mathbf{z}_{is} = 0$ for the away team (and all

other teams). Therefore, considering z_{ir} together with a global effect as in model (10.5) or with a team-specific effect as in model (10.6) corresponds to either considering a global or team-specific home advantages. Object-specific parameters instead of global parameters strongly increase the complexity of the model. Therefore, in contrast to the case of global effects as in model (10.5) regularization seems necessary, as done for the team-specific home advantages in Chapter 8.

Paired Comparisons Outside the Bradley-Terry Framework

In Chapter 9, a paired comparison model outside the Bradley-Terry framework is considered. For data from several FIFA World cups in football, a Poisson model for the number of goals scored by a team against a specific opponent is fitted. Therefore, for every paired comparison (for every match) two outcomes are considered, namely the scores of both teams. The model can be considered to be a paired comparison model as the linear predictor contains the difference between traits of both competing teams. More precisely, the difference between the attack ability of one team and the defense ability of the other team is considered. Compared to the Bradley-Terry model this model has two main advantages for the modelling of football tournaments: As the purpose of the application lies in the prediction of the FIFA world cup, it is necessary to predict exact match outcomes (i.e. the scores of both teams) instead of only distinguishing between wins, draws or losses. Otherwise, an appropriate prediction of the group stage is not possible. Second, it is easier to distinguish between the two important abilities in football, namely attack and defense. The application in Chapter 9 showed that also with this model the inclusion of covariates is reasonable.

Conclusion

In conclusion, this thesis proposes a variety of extensions to existing models for item response and paired comparison models. It is demonstrated that the proposed extensions can be very useful for the statistical analysis of item response and paired comparison data. Furthermore it is demonstrated that regularization techniques provide valuable tools for estimation and a better interpretability of the proposed models. Nevertheless, there are many aspects left that require further attention and should be the target of future research. I hope, this thesis can provide a contribution to some unanswered questions within the topic of modelling item response and paired comparison data.

In the firm believe that Banksy's theory also holds for dissertations:

*I have a theory that you can make any
sentence seem profound by writing the name
of a dead philosopher at the end of it.*

Plato

Appendices

A. Visualization of Categorical Response Models

A.1. Introduction

In Chapters 4, 5 and 7, so-called effect stars were used to visualize the parameter estimates of the proposed methods. Originally, effects stars were proposed to visualize parameter estimates in multinomial or ordered logit models. After all, they are more generally applicable for all kinds of models with a certain group structure within the parameter estimates. In the context of DIF detection with the DIF model (4.2) as proposed in Chapters 4 and 5, all parameters corresponding to one item form a group of parameters. Therefore, all parameters corresponding to one item can be visualized in one effect star. In chapter 7, the parameters can be grouped by covariates and one star represents all parameters corresponding to one covariate. For a better understanding of the basics and the interpretation of effect stars, in the following the principle concept of effect stars in categorical response models is outlined in detail.

Multinomial response models are a common tool in categorical data analysis with well-established theory. But in applications, in particular in the case of many response categories, it is often tedious to keep track and interpret all of the parameters. Therefore tools for visualization of the effects of explanatory variables will be helpful for practioners.

In multivariate data analysis visualization techniques have a long tradition. Skillfully devised graphical methods allow one to look into data and uncover features of the underlying data generating process. They are used to explore data and also to present results. Various books and articles are devoted to graphical representations of data, see, in particular, Cleveland (1985), Kastellec and Leoni (2007), and the Handbook of Data Visualization (Chen et al., 2008).

This chapter is a modified version of Tutz and Schauburger (2013), previous work on the issue can be found in the technical report 117 (Tutz and Schauburger, 2012b) and the conference paper Schauburger and Tutz (2012). See Chapter 1 for more information on the personal contributions of all authors and textual matches.

In the following the focus is not on visualization of data but on the visualization of fitted models to help in the interpretation of parameters. The aim of closer linkage of statistical modelling with graphics is investigated in the case of categorical response models. Categorical response models like the multinomial logit models represent a challenge if the number of response categories and/or the number of explanatory variables is large. Even for moderate numbers of explanatory variables one obtains a large number of parameters and the impact of the predictors on the response variable is hard to investigate because of the transformation to logits. While the increase or decrease of the mean response is easily seen in linear models, the effect on logits is much harder to explain to practitioners.

There has been some work in the visualization of categorical data. In particular graphical methods for the analysis of multiway contingency tables in the form of mosaic plots (Friendly, 1994; Theus and Lauer, 1999; Meyer et al., 2008) are widely used. But categorical response models that also contain continuous predictors cannot be reduced to contingency tables without loss of information. Therefore, for the general case of categorical responses mosaic plots are not very helpful. More recently, in Fox and Andersen (2006) and Fox and Hong (2009) the work on effect displays for generalized linear models (Fox, 2003) was extended to multinomial and proportional-odds logit models, available in the `effects` package (Fox, 2003; Fox and Hong, 2009). The proposed effect displays depict fitted category probabilities including pointwise confidence envelopes and are typically used for visualization of high-order terms. The package provides several kinds of displays for polytomous logit models.

The objective of this appendix is to develop alternative graphical methods for the general case of categorical response models with all types of regressors. In Section A.2 we briefly sketch the multinomial logit model and the interpretation of parameters. In Section A.3 more traditional tools for the graphical representation of the effect of explanatory variables in the form of probability plots are considered. The main tool, graphical tools for the visualization of parameters, is given in Section A.4. We conclude with an extension to ordinal response models.

A.2. The Multinomial Logit Model

In the following we shortly summarize the essential properties of the multinomial logit model, which is the most frequently used model in regression analysis for un-ordered categorical responses and is extensively treated, for example, in Agresti (2002). For response $Y \in \{1, \dots, k\}$ and the vector of explanatory variables \mathbf{x} it has the form

$$P(Y = r|\mathbf{x}) = \frac{\exp(\beta_{r0} + \mathbf{x}^T \boldsymbol{\beta}_r)}{\sum_{s=1}^k \exp(\beta_{s0} + \mathbf{x}^T \boldsymbol{\beta}_s)}, \quad (\text{A.1})$$

where $\boldsymbol{\beta}_r^T = (\beta_{r1}, \dots, \beta_{rp})$. Since parameters $\beta_{10}, \dots, \beta_{k0}, \boldsymbol{\beta}_1^T, \dots, \boldsymbol{\beta}_k^T$ are not identifiable additional constraints are needed. One option is to choose one of the response categories as reference category. For example, by setting $\beta_{k0} = 0, \boldsymbol{\beta}_k = \mathbf{0}$, category k is chosen as the reference category and interpretation of all parameters refers to this category. Alternatively one can use the symmetric side constraints $\sum_{s=1}^k \beta_{s0} = 0, \sum_{s=1}^k \boldsymbol{\beta}_s^T = (0, \dots, 0)$. In both cases one has $k-1$ intercepts and $p(k-1)$ effects of predictors, where p denotes the length of \mathbf{x} . Even for moderate number of predictors, say 10, and 5 response categories, one obtains 40 parameters that represent effects of predictors. The result is a lengthy list of parameter estimates that contains the relevant information but it takes some skill and time to evaluate the effects.

The large number of parameters is due to the multi-dimensionality of the model. The response variable $Y \in \{1, \dots, k\}$ hides the fact that the response is actually multivariate. This becomes obvious by considering the distribution of the response. By defining dummy variables y_1, \dots, y_{k-1} with $Y = r \Leftrightarrow y_r = 1$ the possible outcome vectors of length $k-1$ are given by $(1, 0, \dots), (0, 1, 0, \dots) \dots (0, 0, \dots, 0)$. With probabilities given by $\pi_r(\mathbf{x}) = P(Y = r|\mathbf{x}) = P(y_r = 1|\mathbf{x})$ the vector $\mathbf{y}^T = (y_1, \dots, y_{k-1})$ follows a multinomial distribution $\mathbf{y} \sim M(1, \boldsymbol{\pi}(\mathbf{x}))$, where $\boldsymbol{\pi}^T(\mathbf{x}) = (\pi_1(\mathbf{x}), \dots, \pi_{k-1}(\mathbf{x}))$ represents the vector of response probabilities. A closed representation of the $(k-1)$ -dimensional model as a multivariate generalized linear model (GLM) uses the form $g(\boldsymbol{\pi}(\mathbf{x})) = \mathbf{X}\boldsymbol{\beta}$ with $(k-1)$ -dimensional link function g , design matrix \mathbf{X} and all the parameters collected in $\boldsymbol{\beta}^T = (\beta_{10}, \dots, \beta_{k-1,0}, \boldsymbol{\beta}_1^T, \dots, \boldsymbol{\beta}_{k-1}^T)$. Maximum likelihood estimation and parameter tests can be derived within the framework of multivariate GLMs (see, for example, Tutz, 2012).

For the interpretation of the parameters it is essential to specify the identifiability constraint that is used. If k is chosen as the reference category one obtains

$$\log \left(\frac{P(Y = r|\mathbf{x})}{P(Y = k|\mathbf{x})} \right) = \beta_{r0} + \mathbf{x}^T \boldsymbol{\beta}_r, \quad r = 1, \dots, k-1, \quad (\text{A.2})$$

where the log-odds compare $P(Y = r|\mathbf{x})$ to the probability $P(Y = k|\mathbf{x})$. Then the parameters reflect the effect of predictors on the relation between category r and the reference category k . Symmetric side constraints are less often used although there is a nice interpretation of parameters. For symmetric side constraints the interpretation refers to the "mean" response defined by the geometric mean

$$GM(\mathbf{x}) = \sqrt[k]{\prod_{s=1}^k P(Y = s|\mathbf{x})} = \left(\prod_{s=1}^k P(Y = s|\mathbf{x}) \right)^{1/k}.$$

It is easily derived that

$$\log \left(\frac{P(Y = r|\mathbf{x})}{GM(\mathbf{x})} \right) = \beta_{r0} + \mathbf{x}^T \boldsymbol{\beta}_r, \quad r = 1, \dots, k,$$

holds. Therefore, $\boldsymbol{\beta}_r$ reflects the effects of \mathbf{x} on the logits when $P(Y = r|\mathbf{x})$ is compared to the geometric mean response $GM(\mathbf{x})$.

When visualizing effects we will focus on symmetric side constraints because effects do not refer to the assigned reference category but to all of the categories. Also the results of testing hypotheses and corresponding p -values are easier to interpret. If $H_0 : \beta_{rj} = 0$ is rejected for the model with reference category k the j th variable distinguishes significantly between response $Y = r$ and $Y = k$. If $H_0 : \beta_{rj} = 0$ is rejected for the model with symmetric side constraint the j th variable distinguishes between response $Y = r$ and $Y \neq r$.

A.3. Traditional Methods of Visualization: Probability Plots

When visualizing the effects of predictors the main problem with the multinomial logit model is that the link function is not linear. Although odds are an intuitive concept, the log-odds in equation (A.2) are not appropriate to obtain some feeling for the impact of predictors. Therefore, a traditional way to visualize the effect of explanatory variables is the plotting of response probabilities against the values of specific covariates, see, for example, Agresti (2002).

For illustration we will consider the modelling of party choice with data from the German Longitudinal Election Study. The response categories refer to the dominant parties in Germany, in particular, the Christian Democratic Union (CDU: 1), the Social Democratic Party (SPD: 2), the Liberal Party (FDP: 3), the Green Party (4) and the Left Party (Die Linke: 5). With the five response categories nine predictors were collected, age in years, political interest (1: less interested, 0: very interested), religion (1: evangelical, 2: catholic, 3: otherwise), regional provenance (west; 1: former West Germany, 0: otherwise), gender (1: male, 0: female), union (1: member of a union, 0: otherwise), satisfaction with the functioning of democracy (democracy; 1: not satisfied, 0: satisfied), unemployment (1: currently unemployed, 0: otherwise), and high school degree (1: yes, 0: no).

Table A.1 shows the estimated parameters together with standard errors. It is seen that even in this simple example with moderate number of predictors and response categories many parameters have to be investigated. A simple way to illustrate the effect of a metric

Table A.1.: Estimates of multinomial logit model for party preference data, symmetric side constraints.

	Intercept	Age	Religion (2)	Religion (3)	Democracy	Pol.Interest
CDU	1.397	0.308	0.404	-0.358	-0.766	0.202
SPD	0.469	0.148	-0.196	-0.428	-0.360	0.337
FDP	-0.345	-0.111	0.090	0.326	0.002	-0.264
Greens	-1.096	-0.398	-0.127	0.286	0.008	0.214
Left Party	-0.425	0.053	-0.171	0.174	1.116	-0.488

	Unemployment	Highschool	Union	West	Gender
CDU	-0.514	0.156	-0.408	-0.330	-0.262
SPD	0.127	-0.221	0.400	0.389	-0.191
FDP	-0.560	0.051	-0.509	0.025	0.254
Greens	-0.071	0.563	-0.391	0.639	-0.019
Left Party	1.018	-0.549	0.907	-0.723	0.218

Standard Errors

	Intercept	Age	Religion (2)	Religion (3)	Democracy	Pol.Interest
CDU	0.224	0.069	0.163	0.168	0.139	0.147
SPD	0.245	0.072	0.166	0.172	0.148	0.160
FDP	0.312	0.094	0.239	0.218	0.194	0.191
Greens	0.327	0.097	0.234	0.213	0.193	0.203
Left Party	0.313	0.094	0.233	0.205	0.229	0.185

	Unemployment	Highschool	Union	West	Gender
CDU	0.366	0.169	0.212	0.156	0.135
SPD	0.314	0.189	0.184	0.172	0.142
FDP	0.498	0.218	0.289	0.207	0.187
Greens	0.421	0.208	0.273	0.222	0.184
Left Party	0.301	0.243	0.216	0.194	0.181

covariate like age is to plot the response probabilities against age. But, of course, in a non-linear model as the logit model, the form of the function strongly depends on the values of the other parameters. In Figure A.1 the probabilities are given for two sets of values, one where all other predictors have value 0, one where all other predictors have value 1. It is seen that not only the level but also the slope of the curves can vary with the chosen value for the other variables. For example, the curve for the Social Democratic Party (SPD) is rather flat in the upper panel, but increasing in the lower panel. When explanatory variables are categorical, bar plots with the probabilities corresponding to the height of the bars can be used. Figure A.2 shows the effect of unemployment on the choice probabilities. It shows, for example, that unemployed persons have a stronger preference for the left party, preference for CDU decreases. The tendency is the same if different values are chosen for the other variables (Figure A.3), but effect strength is quite different. If the other variables

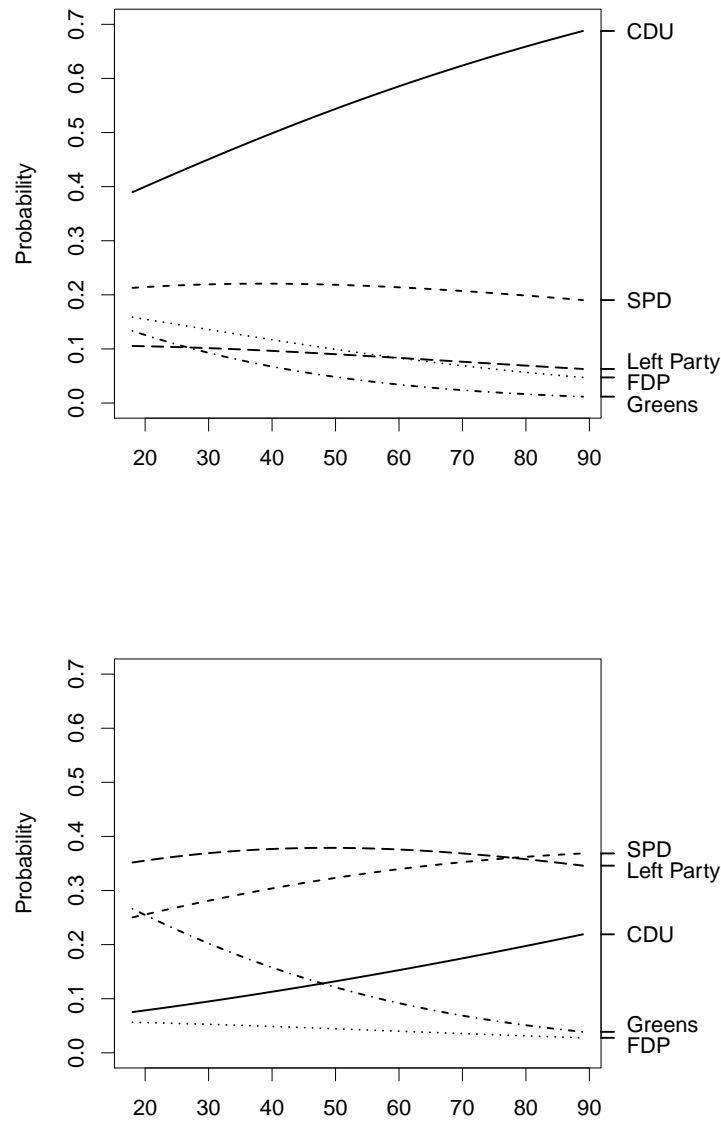


Figure A.1.: Estimated probabilities for party preference against age, all other variables fixed at value 0 (upper panel), all other variables fixed at value 1 (lower panel).

have value 1, the probability for CDU is among the lowest if voters are unemployed. Thus the values of the unplots variables can and do make a difference in the response profiles for the predictor variable which is plotted.

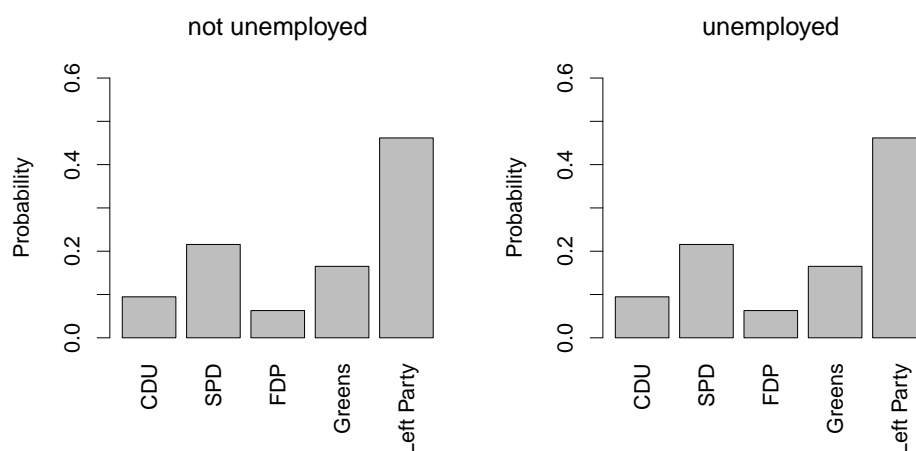


Figure A.2.: Bar plot of estimated probabilities for party preference for unemployment=0 (left) and unemployment=1 (right), all other categorical variables fixed at value 0, age at 50.

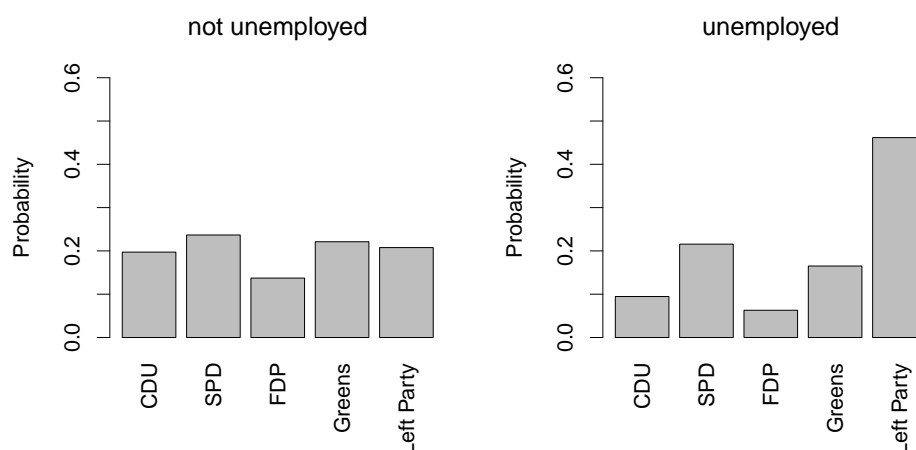


Figure A.3.: Bar plot of estimated probabilities for party preference for unemployment=0 (left) and unemployment=1 (right), all other categorical variables fixed at value 1, age at 50.

A.4. Glyphs for the Visualization of Parameters

The disadvantage of bar plots as well as curves is that they show effects under the constraint that the other predictors have fixed chosen values. The plots vary with the chosen values. An alternative approach that is propagated here is to visualize the effect strength that is contained in the parameters rather than the probabilities.

We will use glyphs that have traditionally been used to visualize data. Various glyphs have been proposed in the literature, among them profile glyphs (Du Toit et al., 1986), Chernoff faces (Chernoff, 1973) and stars (Anderson, 1957, Siegel et al., 1972, Gnanadesikan, 1977).

We will focus on star plots, but instead of using them to visualize data, they are used to visualize parameters. The parameters of the logit model themselves are less appropriate since they contain the effect on logits, which do not carry much intuition. A much better way is to focus on the odds that stand behind the log-odds (or logits).

A.4.1. Star Plots for Parameters

The main tool is the representation of the odds of a model with symmetric side constraints as

$$\frac{P(Y = r|\mathbf{x})}{GM(\mathbf{x})} = \exp(\beta_{r0} + \mathbf{x}^T \boldsymbol{\beta}_r) = e^{\beta_{r0}} e^{x_1 \beta_{r1}} \dots e^{x_p \beta_{rp}} = e^{\beta_{r0}} (e^{\beta_{r1}})^{x_1} \dots (e^{\beta_{rp}})^{x_p}.$$

From

$$\frac{P(Y = r|x_1, \dots, x_j + 1, \dots, x_p)/GM(x_1, \dots, x_j + 1, \dots, x_p)}{P(Y = r|x_1, \dots, x_j, \dots, x_p)/GM(x_1, \dots, x_j, \dots, x_p)} = e^{\beta_{rj}}$$

it is seen that $e^{\beta_{rj}}$ represents the multiplicative effect of variable j on the odds $P(Y = r|\mathbf{x})/GM(\mathbf{x})$ if x_j increases by one unit.

In "effect stars", which are proposed here, the lengths of the rays emanating from the center of the plot represent the exponentials of the parameters. Thus one obtains a star plot for each variable that shows how strong the impact of the predictor on the response is and what form it takes. In addition, we include a (shaded) unit circle around the center that corresponds to the no-effects case, where $\beta_{1j} = \dots = \beta_{kj} = 0$ or, equivalently, $e^{\beta_{1j}} = \dots = e^{\beta_{kj}} = 1$ holds. Therefore, the deviation from the circle shows the strength of the preference for one category as the deviation from the circle. If the ray is outside the circle the increase in the predictor increases the probability of the corresponding category, if it is inside the circle the increase in the predictor decreases the response probability. Stars are standardized such that the maximal length of a ray has the same value. This value also scales the radius of the unit circle.

Figure A.4 shows the effect stars for the main effect model fitted to the party choice data, where the quantitative variable age has been standardized. Let us consider the effect of age. It is immediately seen that with increasing age the Christian-democratic party (CDU) is more strongly favored while, in particular, the response probability for the Greens decreases. An additional feature that is included is the significance of the deviation. The value in brackets given at each ray is the p -value of the hypothesis $H_0 : \beta_{rj} = 0$ for the model with symmetric side constraint. The effects of age on responses CDU, SPD and Greens turned out to be significant at the level 0.05, the former two with positive (outside the circle), the latter with negative effect (within the circle). In addition, the overall p -value for the hypothesis that one variable can be neglected, that is, $H_0 : \beta_{1j} = \dots = \beta_{kj} = 0$, is given

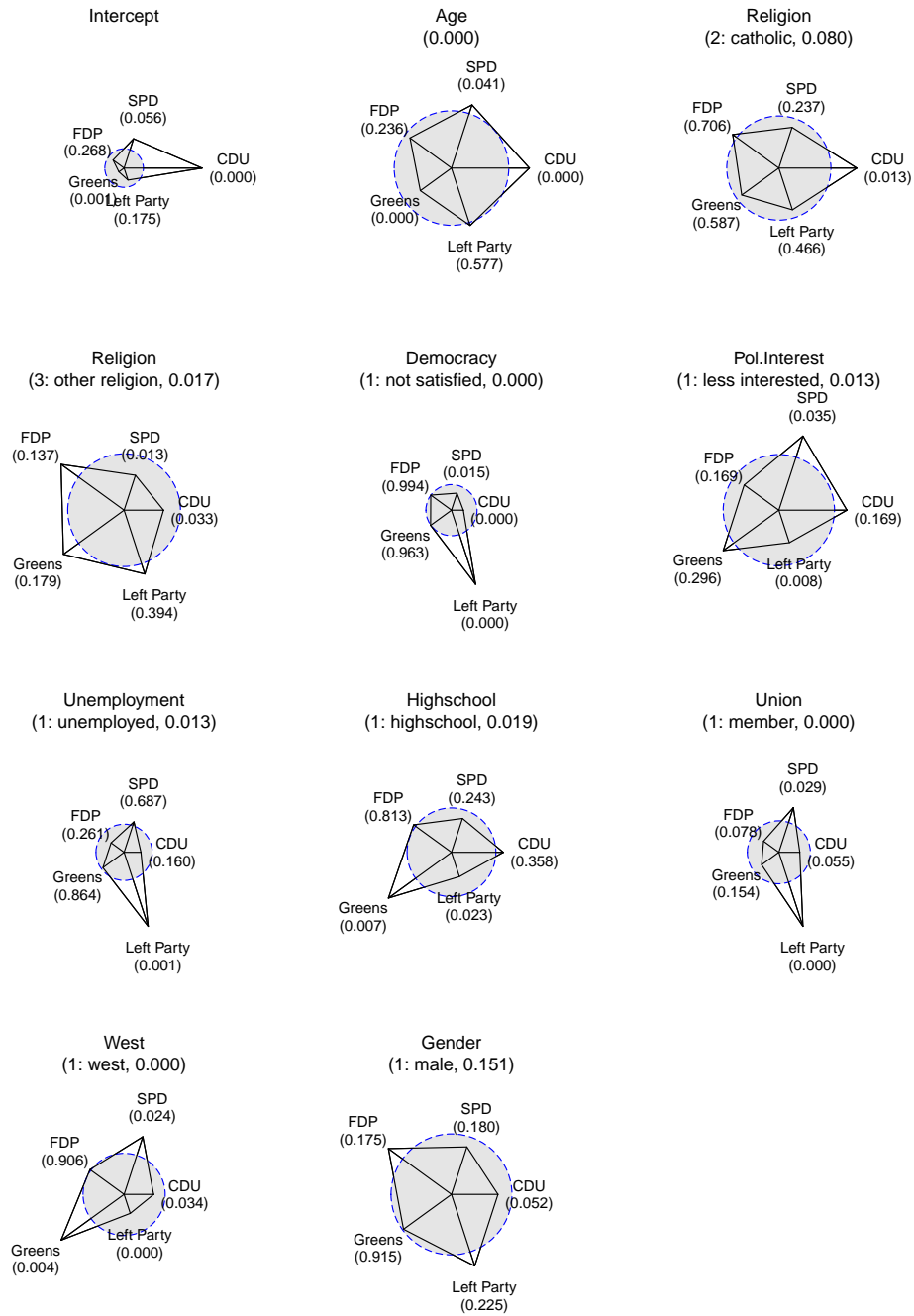


Figure A.4.: Effect stars showing the exponentials of parameters, p -values at the rays refer to hypothesis $H_0 : \beta_{rj} = 0$, p -values given with the variable description refer to hypothesis $H_0 : \beta_{1j} = \dots = \beta_{kj} = 0$.

with the description of the variable. For example, age turned out to be highly significant (0.000), whereas the effect of gender was weak (p -value of 0.151).

The advantage of the effect star plots is that all the effects of the variables are shown simultaneously. Discrete as well as continuous variables are given in the same representation. In addition to the direction of the effect seen from the shape of the star, information on the significance of specific effects is included, as well as information about the whole variable.

Relevant features are easily seen from the shape of the stars. For example, very strong deviations from the circle are found for the variables democracy, unemployment and union. All these variables have a strong effect in favor of the left party. Deviations from the star in favour of the Greens are seen for the variables high school and west. Supporters of the Greens are found among more educated persons from the former west.

A.4.2. Extensions and Alternatives

The presentation can be extended to include standard errors. Let se_{rj} denote the standard error for estimation of β_{rj} . Then, an approximative confidence interval for the exponential is given by $[\exp(\hat{\beta}_{rj} - 1.96se_{rj}), \exp(\hat{\beta}_{rj} + 1.96se_{rj})]$. By plotting the lower and the upper limit one obtains an inner and an outer star.

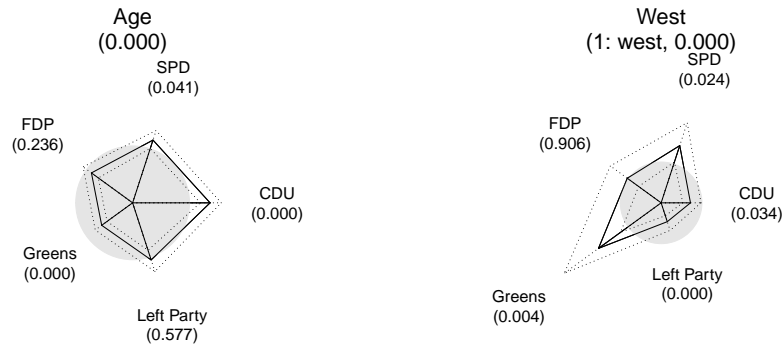


Figure A.5.: Effect stars with reliability intervals for two variables (party preference data)

Figure A.5 shows the plots of two predictors for the party preference data. If p -values are large, for example, for FDP and the left party for variable age, and FDP for variable west, the circle is covered by the corresponding intervals whereas for highly significant predictors, for example, CDU for variable age, the corresponding intervals are outside or within the circle. Inclusion of standard errors is certainly helpful but with many stars information content can be high. One strategy is to look first at all the stars without reliability intervals and then pick out the interesting ones and look at them more closely.

It should be noted that star plots for the exponentials of the parameters have the same form if a reference category is chosen. But then a more appropriate circle is the circle with radius defined by the reference category. The radius is fixed by the length of the ray for the reference category. Figure A.6 shows effects of two variables with reference category CDU. Now rays inside the circle show that the predictor decreases the preference for the corresponding category when compared to the reference category. Rays outside the circle represent the opposite effect. But in both cases interpretation is in relation to the specified reference category (CDU). Consequently the p -values given now refer to the null hypothesis $H_0 : \beta_{rj} = 0$ for parameters constrained by fixing the reference category.

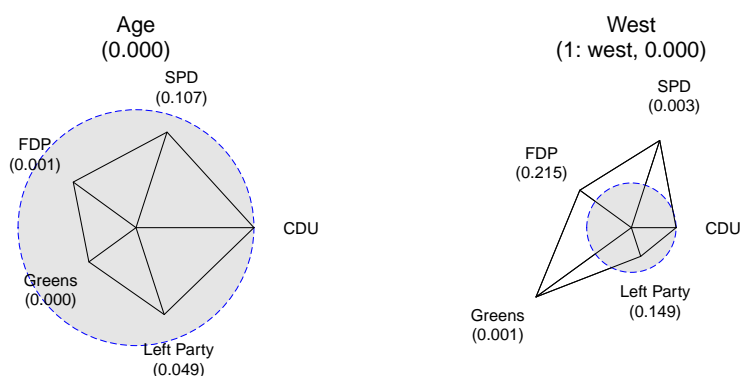


Figure A.6.: Effect stars with reference category (party preference data)

In Figure A.4 the main effect model was represented by star plots. Also interaction terms can be represented as stars but the representation is less useful than for main effect models. The reason is that the effects of variables that are included in an interaction effect are more difficult to interpret. For simplicity we consider one interaction term that turned out to be significant, namely the interaction between age and the binary variable democracy (1: not satisfied, 0: satisfied). Figure A.7 shows the stars for the marginal terms and the interaction. The stars for the other variables hardly change when the interaction is included and therefore are not shown. Compared to Figure A.4 one sees that the main effect of democracy hardly changes while the main effect of age is quite different. Nevertheless, interpretation differs from that of the main effect models. The effect of age now represents the age slope among those who are satisfied and the effect of democracy represents the difference between not satisfied and satisfied at mean age because age was standardized. The interaction effect contains the modification of the effect of one variable by the other. It represents the difference in age slope between not satisfied and satisfied. It is seen that the preference for the big parties, SPD and CDU, increases stronger with age in the not satisfied group than in the satisfied group whereas for the green and the left party the dependence on age is weakened if voters are not satisfied with democracy. Since interpretation is much

harder when interaction effects are included alternative visualization tools as given in the next section are to be recommended.

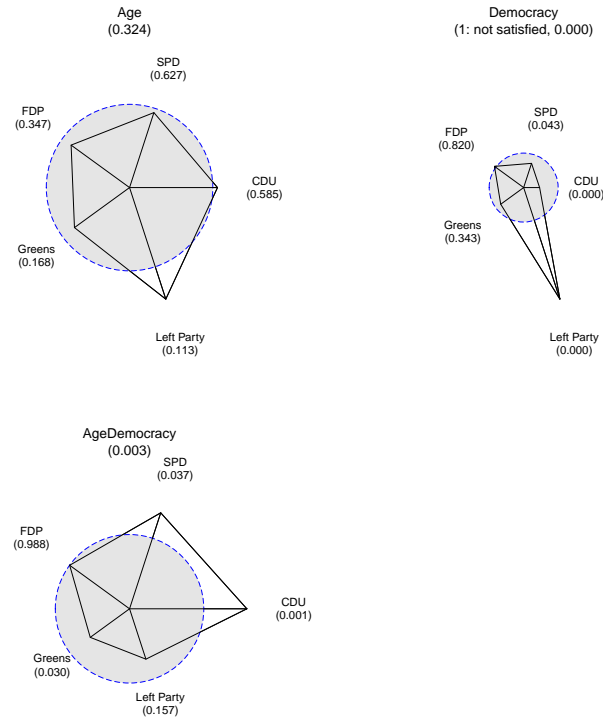


Figure A.7.: Marginal and interaction term stars for standardized age and democracy (1: not satisfied, 0: satisfied) for party preference data.

A.4.3. Alternative Displays

Star plots visualize parameters of fitted models. The plots are especially simple for main effect models when predictors are binary or are measured on a metric scale level. Then one star collects all the parameters connected to one explanatory variable. For categorical predictors with more than two categories several stars are linked to one predictor. The same holds when interactions are included. Then one has at least three stars that are linked to two variables. Although the interaction star as a visualization of the underlying effects is interpretable, the effect of a variable is not easily seen since it has to be seen in combination with the variable with which it interacts. The effect displays proposed by Fox and Andersen (2006) are able to visualize the effects of interaction terms quite nicely by allowing other predictors marginal to a given term to be set at average or other values.

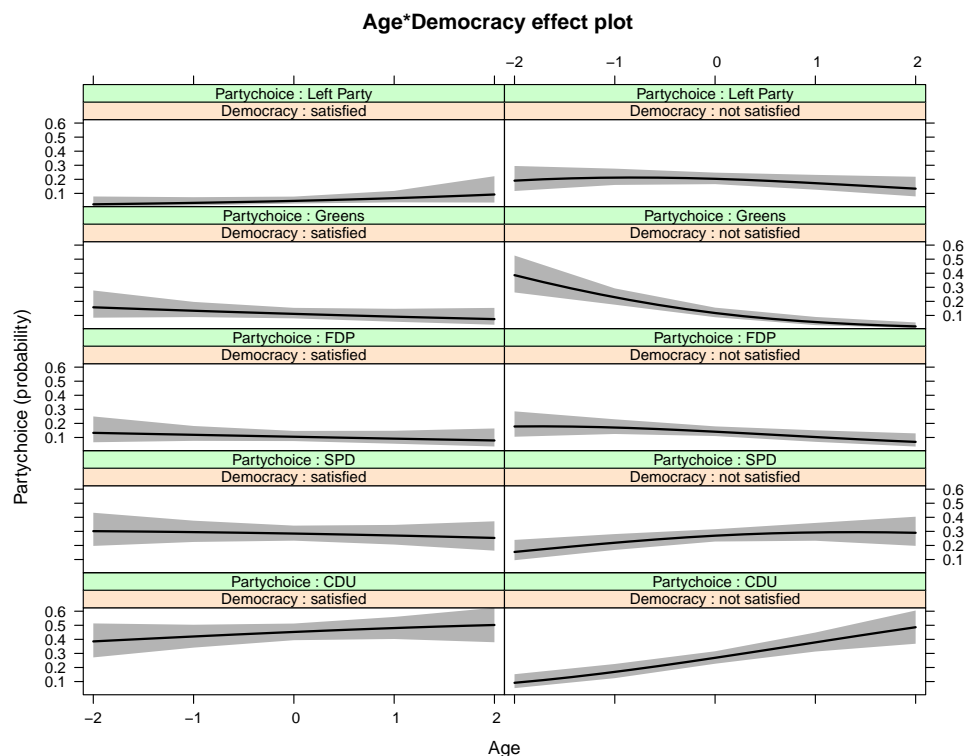


Figure A.8.: Effect plot for the interaction of age and democracy

For illustration we consider the interaction effect between the variables age and democracy, for which stars are given in Figures A.8 and A.9. Figure A.8 shows the typical effect plots as curves and Figure A.9 shows the "stacked area" displays also offered by the `effects` package (Fox, 2003; Fox and Hong, 2009). They visualize what the interaction star in Figure A.7 shows only qualitatively, that the preference for the big parties, SPD and CDU, increases stronger with age if voters are not satisfied, for the green and the left party the effect slope decreases if voters are not satisfied. In particular the stacked area display visualizes nicely the effect of age and democracy on the response. Nevertheless, it should be noted that the effects on the probabilities are shown for fixed values of the other variables, in our case they have been chosen by mean values. If other values are chosen the effects on probabilities might change.

One can also plot the linear predictor itself, which means the effect on the logits. This plots would essentially show the same form of effects (but shifted) for other values of the rest of the variables, but it has the disadvantage that it is much harder to think in logits than in probabilities. For binary responses the `effects` package offers the option to label the response axis nonlinearly on the probability scale. Then one can see the effect on probabilities from the scaling. For multinomial the scaling is not so straightforward because it depends on the logits that were chosen, that is, the reference category that has been fixed. In their application Fox and Hong (2009) also rely on probability plots

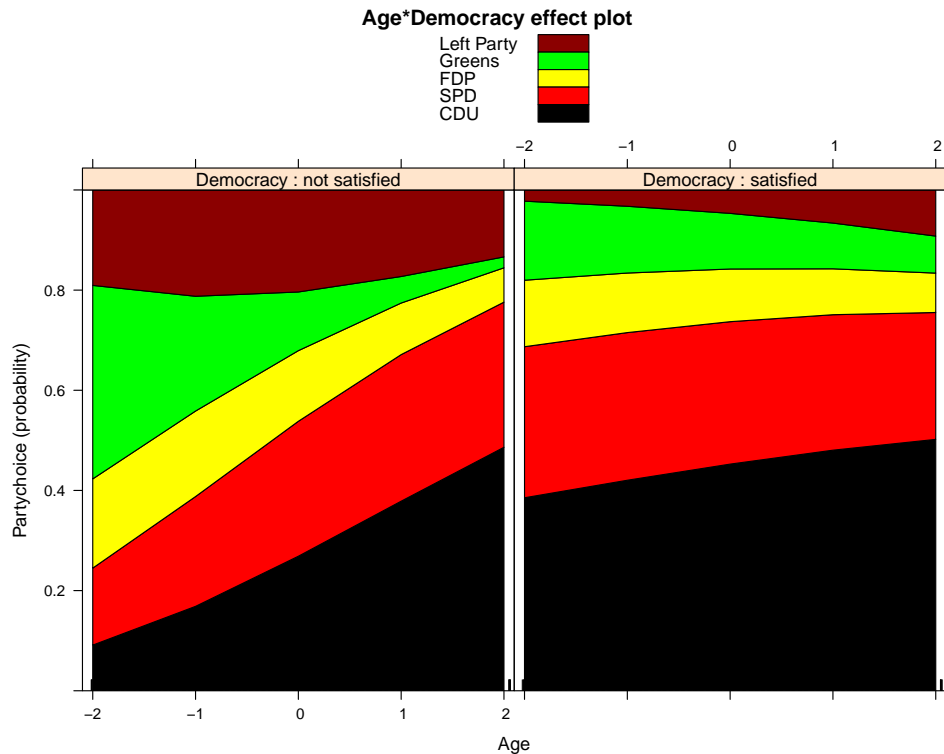


Figure A.9.: Stacked area display for the interaction of age and democracy.

to visualize the effects in multinomial models. Star plots avoid the dependence on the reference category by using symmetric side constraints. By using odds rather than logits the effect strength is somewhat more intuitive.

The essential difference between stars and effect displays provided by the **effects** package is that stars visualize parameters with effect strength referring to specific odds and effect displays visualize the effects on probabilities or logits as curves. Effects displays are strong tools especially for interaction effects because they include the marginal effects. After screening the effects by star plots it is certainly a good idea to look at the plots provided the **effects** package, which, in particular for metric predictors, show the continuous dependence on the predictor. One other advantage of the **effects** package is that smooth effects of continuous predictors can be included. Although one might construct stars that visualize smooth effects it would destroy the simplicity of the visualization by stars (see concluding remarks).

A.4.4. Further Examples

For further illustration we consider brand choice data. The data refer to different brands of coffee. The purchases of coffees of 2111 households were collected by the

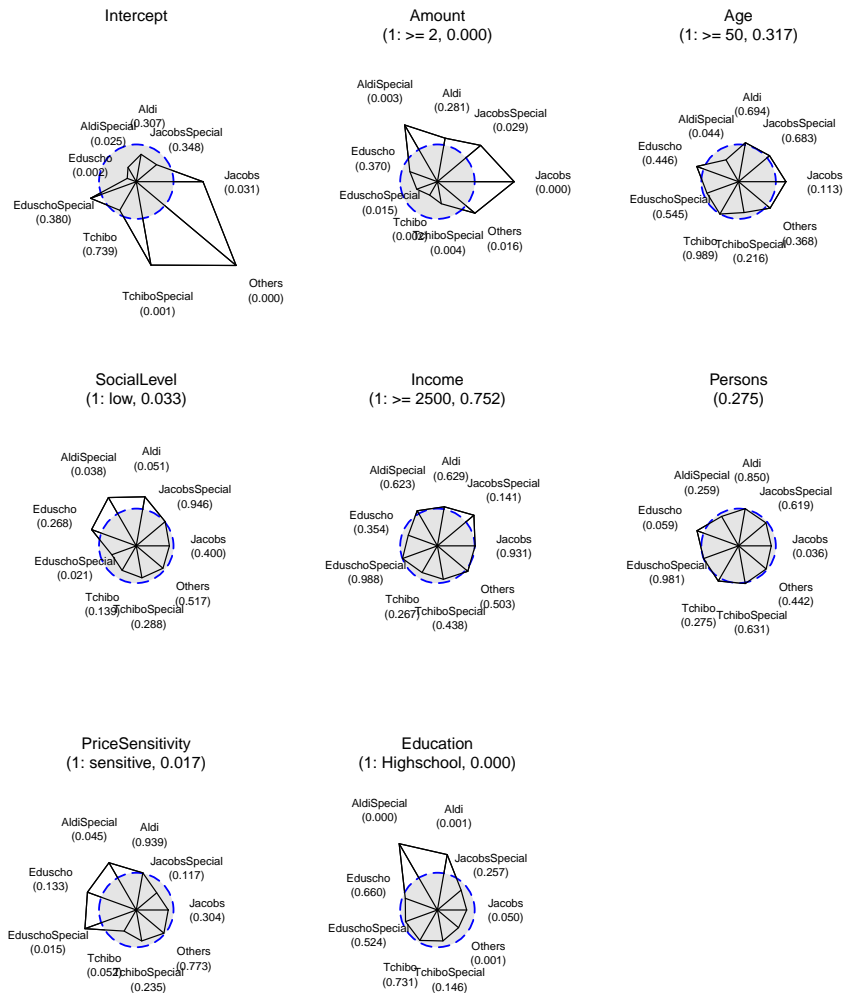


Figure A.10.: Graphs for brand choice data with seven predictors and fixed radius. It is seen that stars for the significant predictors, mount, social level, price sensitivity and education, deviate strongly from the circle.

Gesellschaft für Konsumforschung (Society for Consumer Research) and are available at <http://www.statistik.lmu.de/service/datenarchiv/kaffee/kaffee.html> or in the R-package *EffectStars* (Schauberger, 2014b). The brands were named after the shops, which offer a regular brand and a special brand, Aldi, AldiSpecial, Jacobs, JacobsSpecial, Eduscho, EduschoSpecial, Tchibo, TchiboSpecial. The binary covariates were the number of packages bought (amount; 1: ≥ 2), age (1: ≥ 50), social level (1: low), monthly income (1: ≥ 2500), persons in household, price sensitivity (1: sensitive), education (1: high school). Figure A.10 shows the corresponding glyphs. Three of the predictors are not significant,

namely age, income, and persons in household. It is seen that the corresponding stars are very close to the circle. For the significant predictors the stars deviate strongly from the no-effects circle. Naturally, the interpretation of the single effects refers to the brands considered. One sees, for example, that the brands, offered by the cheap discounter Aldi, are preferred if the social level is low. The stars are scaled in a different way, namely by fixing the radius of the unit circle. What works well in this example can be less advantageous for other data (see next example).

An often used example with a categorical predictor is the alligator food choice considered in Agresti (2002). In the study by the Florida Game and Fresh Water Commission the response is the primary food type in categories fish, invertebrate, reptile, bird, and other. The explanatory variables are size, dichotomized into $\leq 2.3, > 2.3$, gender (1: male, 0: female), and the lake where the reptiles lived (four categories, 1: George, 2: Hancock, 3: Oklawaha, 4: Trafford), see Agresti (2002). A problem with categorical variables like the lake is that a reference category has to be chosen. This can be avoided by using effect coding for the predictor by using a symmetric side constraint. Let the categorical variable A have values $1, \dots, m$, and $\beta_{r,A(j)}$ denote the parameter for category j of the predictor and response category r . Then the symmetric side constraint is given by $\sum_{j=1}^m \beta_{r,A(j)} = 0$. Interpretation of parameters with the symmetric side constraint does not refer to an increase by one unit but always refers to a mean over categories. Let $GM(A = j, \mathbf{x}_R)$ denote the geometric mean defined in Section 2 with the predictor A being in category j and the rest of the variables represented by \mathbf{x}_R . With this notation one derives for the multinomial model

$$e^{\beta_{r,A(j)}} = \frac{P(Y = r|A = j, \mathbf{x}_R)/GM(A = j, \mathbf{x}_R)}{\prod_{s=1}^m P(Y = r|A = s, \mathbf{x}_R)/GM(A = s, \mathbf{x}_R)},$$

which compares the odds for predictor value $A = j$, $P(Y = r|A = j, \mathbf{x}_R)/GM(A = j, \mathbf{x}_R)$, to the geometric mean response probability over all categories of variable A for fixed \mathbf{x}_R .

For effect coded predictors one gets as many stars as categories whereas one has one less if a reference category is chosen (see religion in the party preference example). Figure A.11 shows the resulting glyphs with effect coding for the lakes. It is seen that size of the alligator changes the food preference; larger alligators have a stronger preference of bird and reptiles. Also the lake makes a difference showing that different food is preferred or available in the lakes. Here the advantage of the symmetric side constraint is that preference has not to be interpreted with respect to an arbitrarily chosen reference lake. For illustration in this example we hold the radius constant instead of the maximal length of rays. The option to fix the radius is included in the **EffectStars** package.

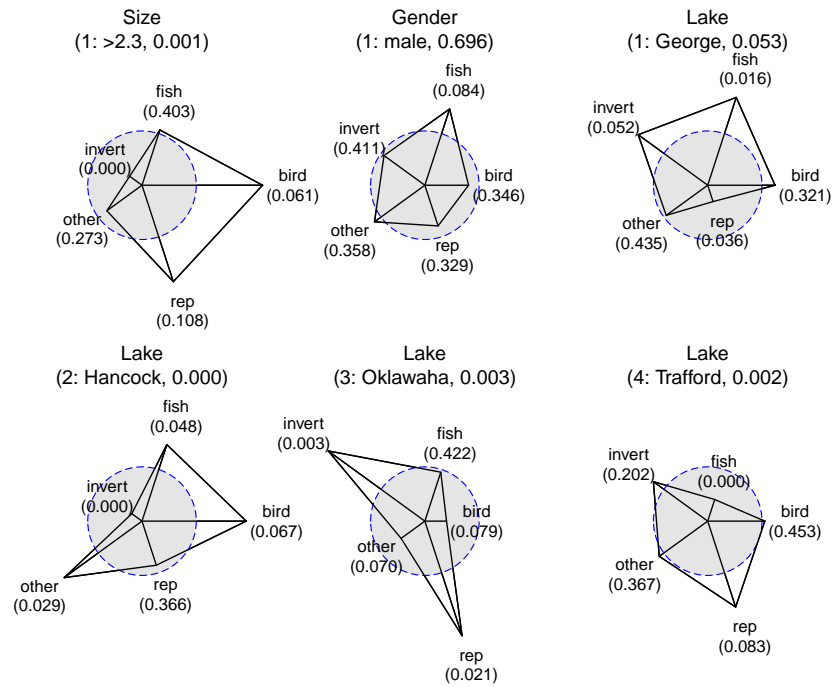


Figure A.11.: Food choice for alligator data depending on size gender and lake with fixed radius.

A.5. Ordinal Response Models

The graphical tool of parameter glyphs can also be used to uncover structures in ordinal response models as the cumulative type models or the sequential type models (for example, Agresti, 2009). For simplicity, we restrict consideration to logit models. Let the response Y take values from ordered categories $\{1, \dots, k\}$. The cumulative logit model has the general form

$$\log \left(\frac{P(Y \leq r | \mathbf{x})}{P(Y > r | \mathbf{x})} \right) = \gamma_{0r} + \mathbf{x}^T \boldsymbol{\gamma}_r, \quad r = 1, \dots, k-1,$$

or

$$P(Y \leq r | \mathbf{x}) = \frac{\exp(\gamma_{r0} + \mathbf{x}^T \boldsymbol{\gamma}_r)}{1 + \exp(\gamma_{r0} + \mathbf{x}^T \boldsymbol{\gamma}_r)}, \quad r = 1, \dots, k-1,$$

The sequential logit model (also called continuation ratio model) has the form

$$\log \left(\frac{P(Y = r | \mathbf{x})}{P(Y > r | \mathbf{x})} \right) = \gamma_{0r} + \mathbf{x}^T \boldsymbol{\gamma}_r, \quad r = 1, \dots, k-1,$$

or

$$P(Y = r | Y \geq r, \mathbf{x}) = \frac{\exp(\gamma_{r0} + \mathbf{x}^T \boldsymbol{\gamma}_r)}{1 + \exp(\gamma_{r0} + \mathbf{x}^T \boldsymbol{\gamma}_r)}, \quad r = 1, \dots, k-1.$$

The model is strongly related to discrete hazard models if the response refers to categorical survival. Then the probability $P(Y = r|Y \geq r, \mathbf{x})$ represents the probability of failure in (time) category r given category r is reached, which is a discrete hazard. For details see, for example, Tutz (2012).

In both models the predictor has the form $\eta_r = \gamma_{0r} + \mathbf{x}^T \boldsymbol{\gamma}_r$. By allowing for *category-specific* effects $\boldsymbol{\gamma}_r^T = (\gamma_{r1}, \dots, \gamma_{rp})$ the model has as many parameters as the multinomial logit model. In its simpler version, where $\boldsymbol{\gamma}_r = \dots = \boldsymbol{\gamma}_{k-1} = \boldsymbol{\gamma}$ holds, the cumulative type model is also called the proportional odds model. Only in this form does it fully use the ordering of the response categories. An intermediate case, where only some of the parameters are category-specific is the partial proportional odds model (for example, Cox, 1995, Brant, 1990, Peterson and Harrell, 1990). With many predictors it is a demanding problem to find out which parameters can be specified as *global*, that is, not varying over categories, and which ones as category-specific. In the exploration of the general model star plots can be helpful.

For the representation of effects it is useful to represent the models in a slightly different form. The cumulative logit model can be written as

$$\frac{P(Y \leq r|\mathbf{x})}{P(Y > r|\mathbf{x})} = e^{\gamma_{r0}} e^{x_1 \gamma_{r1}} \dots e^{x_p \gamma_{rp}} = e^{\gamma_{r0}} (e^{\gamma_{r1}})^{x_1} \dots (e^{\gamma_{rp}})^{x_p}.$$

Therefore, the exponential $e^{\gamma_{rj}}$ represents the multiplicative effect of variable j on the cumulative odds $P(Y \leq r|\mathbf{x})/P(Y > r|\mathbf{x})$ if x_j increases by one unit. It is the effect on the dichotomization into response categories $\{1, \dots, r\}$ and $\{r+1, \dots, k\}$. For the sequential logit model one obtains

$$\frac{P(Y = r|\mathbf{x})}{P(Y > r|\mathbf{x})} = e^{\gamma_{r0}} e^{x_1 \gamma_{r1}} \dots e^{x_p \gamma_{rp}} = e^{\gamma_{r0}} (e^{\gamma_{r1}})^{x_1} \dots (e^{\gamma_{rp}})^{x_p}.$$

Therefore, the exponential $e^{\gamma_{rj}}$ represents the multiplicative effect of variable j on the continuation ratio odds $P(Y = r|\mathbf{x})/P(Y > r|\mathbf{x})$ if x_j increases by one unit.

In a star plot for the effects of variable x_j the length of the rays is given by $e^{\gamma_{1j}}, \dots, e^{\gamma_{k-1,j}}$. As in the multinomial logit model the (dashed) unit circle refers to the case where the j th variable can be neglected, that is, $\gamma_{1j} = \dots = \gamma_{k-1,j} = 0$. The p -value of the likelihood ratio test for the corresponding hypothesis $H_0 : \gamma_{1j} = \dots = \gamma_{k-1,j} = 0$ is denoted by $p\text{-rel}$ since the relevance of the j th predictor is tested. When compared to the circle the stars show if the effects are larger than 1 (outside the circle) or smaller than 1 (inside the circle). In the sequential model that means that a variable that has values within the circle decreases the odds $P(Y = r|\mathbf{x})/P(Y > r|\mathbf{x})$, rays outside the circle represent variables that increase the odds. The interpretation of stars is the same as for the multinomial model, that is, closeness to the unit circle means that the variable is not influential.

For illustration we consider the data from the German Munich founder study. Data were collected on business founders who registered their new companies at the local chambers of commerce in Munich and surrounding administrative districts. The focus was on survival of firms measured in 7 categories, the first six represent failure in intervals of six months, the last category represents survival beyond 36 months. Various covariates are available, economic sector (1: industry, manufacturing companies and building sector, 2: commerce, 3: service industry), legal form (1: small trade without entry in the register of companies, 2: one man business merchant, 3: GmbH, GmbH & CoKG, 4: GbR, KG, OHG), location (0: residential area, 1: business area, industrial area or mixed), new (0: new foundation, 1: partial take-over, take-over, miscellaneous), pecuniary reward (0: main occupation, 1: additional occupation), seed capital (1: > 25000 , 0: ≤ 25000), equity capital (1: yes, 0: no), debt capital (1: yes, 0: no), market (0: local market, 1: national market), clientele (0: wide spread, 1: small amount of important customers), education of founder (1: A-levels, 0: minor), gender of founder (1: male, 0: female), experience (1: > 10 years, 0: ≤ 10), number of employees (1: > 2 , 0: ≤ 2), age of founder. The data of the Munich founder study have also been used by Brüderl et al. (1992) and Kauermann et al. (2005) and are available from the Central Archive for Empirical Social Research, University of Cologne, Germany (<http://www.gesis.org/en/institute/>). We restrict our analysis to those firms that were founded completely new, which leaves us with 1224 cases. We fitted the full sequential logit model with all 18 predictors but show only four of the stars that resulted. Figure A.12 shows the stars for sector3, legal3, location, and new foundation. It is seen that the first two variables are highly significant. The variable sector3 has all values outside the circle, meaning that the odds increase if the firm is in the service industry as compared to reference category 1 (industry). For variable legal3 the star is distinctly inside the circle meaning that legal form 3 decreases the odds when compared to reference category 1 (small trade). For the other variables, location and new foundation, the stars are very close to the circle. Consequently both predictors are not significant (see value in brackets).

In ordinal models a second effect is interesting, namely if the effects are category-specific or global, that is, do the effects of variables vary across response categories or not. Therefore, a second (dotted) circle refers to the model with global effects only, that is, $\gamma_{1j} = \dots = \gamma_{k-1,j} = \gamma_j$. We fit the model that contains all predictors with category-specific effects with the exception of predictor j , which has global effect and include the circle with radius $\exp(\gamma_j)$. The interpretation of stars with respect to the dotted circle is different. Closeness to this circle means that the variable is global, strong deviation signals that it is category-specific. The hypothesis $H_0 : \gamma_{1j} = \dots = \gamma_{k-1,j} = \gamma_j$ investigates if the proportional odds assumption holds for the j th predictor. The corresponding p -value of the test is denoted by p -global since the test investigates if the predictor has global effect. Figure A.13 again shows the stars for only some predictors although the full model has been fitted. All of the predictors that are shown have significant effects (p -rel < 0.05). For predictors sector3,

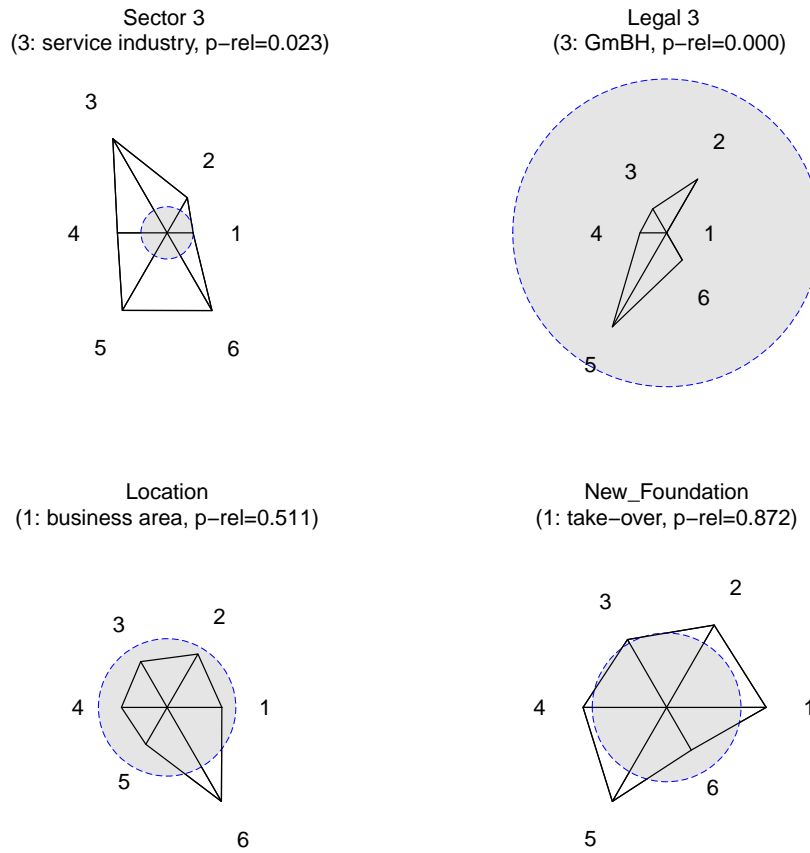


Figure A.12.: Stars for four predictors of the founder study with circles referring to the no-effects case. For the significant variables, sector3 and legal3 the stars are far away from the no-effects circle, for the significant variables, location and new foundation, they are quite close.

legal3 and clientele the hypothesis that effects are global is not rejected ($p\text{-global} > 0.05$). The corresponding stars are close to the dotted circle, although not close to the dashed circle, which represents relevance. For variables legal2 and debt capital the stars are far away from the dotted circle and the hypothesis that the effects are global is rejected. For the latter variable the dashed and the dotted circle are very close, which means that the estimated global effects are very small. However, the variable is influential, but influence becomes relevant only if one allows for category-specific variables. This is one of the cases where variables are excluded if one assumes a model that is too simple but is seen to be relevant if one uses a model that is sufficiently flexible.

In the illustrations we used the sequential model. There are two reasons. First, the category-specific effects for the sequential model have a simple interpretation. Second, the cumulative model often raises problems when a model with category-specific effects is fitted. Maximum likelihood (ML) estimates may not exist because the parameter space is restricted in a

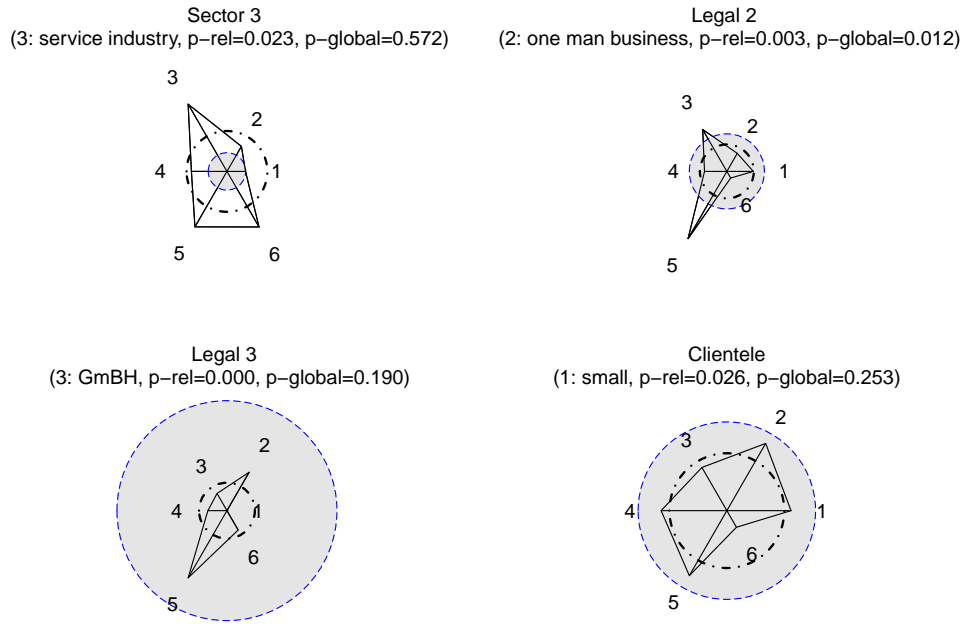


Figure A.13.: Stars for five predictors of the founder study. Deviation from the dashed circle implies relevance of the predictor, deviation from the dotted circle implies that predictor is category-specific.

complicated way, one has to postulate that $\gamma_{10} + \mathbf{x}^T \boldsymbol{\gamma}_1 \leq \dots \leq \gamma_{k-1,0} + \mathbf{x}^T \boldsymbol{\gamma}_{k-1}$ holds for all possible predictor values. If the maximum likelihood estimate does not exist an alternative is to use in the star plot for variable x_j values from the fitting of the global model, which gives the circle, and values from the fitting of the model

$$\log \left(\frac{P(Y \leq r | \mathbf{x})}{P(Y > r | \mathbf{x})} \right) = \gamma_{r0} + x_1 \gamma_1 + \dots + x_j \gamma_{rj} + \dots + x_p \gamma_p,$$

where only variable x_j has category-specific effects. But even then ML estimates often deteriorate.

A.6. Concluding Remarks

We proposed a method to visualize the fitted effects of a categorical response model. The method allows to identify the direction as well as the strength of the effects. For ordinal models it is distinguished between the relevance of a predictor and how strongly the effects vary across the categories. Both aspects can be seen from the corresponding stars. The full strength of the visualization method is seen if one looks at the stars for all the covariates. In

particular in the ordinal response case we showed only selected stars although much more predictors were used.

Star plots visualize parametric effects and therefore are useful for parametric models. They should not be used if a metric variable is included in polynomial form and therefore represented by a group of parameters. Also if multi-category predictors or interaction effects are included one predictor is represented by more than one parameter for each response category and therefore several stars are linked to one predictor. In the case of a multi-category predictor in the typically used parametrization the parameters refer to a chosen reference category and so do the stars. If the predictor has m categories one obtains $m - 1$ stars that have to be interpreted as contrasts to the reference category. If one wants to avoid a reference category one can use a symmetric side constraint and obtains m stars, one for each category of the predictor. For illustration the symmetric side constraint has been used to code the predictor lake in the alligator food example. When interaction effects are included at least two stars are linked to one predictor and some care is needed when interpreting stars because interpretation of parameters is much harder.

From one perspective star plots can be seen as profile plots rendered in polar coordinates. Therefore, profile plots are an alternative, in particular for multinomial models. But we think that stars plots are more pleasing to the eye and the inclusion of the results of significance tests, in particular for ordinal models is easier in star plots.

All the computations were done by use of the free software **R** (R Core Team, 2015). The **R** package **EffectStars** (Schauberger, 2014b) that generates and plots effect stars is available at CRAN. It contains many options to modify the resulting stars.

B. Identifiability of the DIF Model

Proposition

Let for the parameters of the general DIF model (4.2) with predictor $\eta_{pi} = \theta_p - \beta_i - \mathbf{x}_p^T \boldsymbol{\gamma}_i$ be constrained by $\beta_I = 0$, $\boldsymbol{\gamma}_I^T = (0, \dots, 0)$ and let the matrix \mathbf{X} with rows $(1, \mathbf{x}_1^T), \dots, (1, \mathbf{x}_P^T)$ have full rank. Then parameters are identifiable.

Proof

Let two sets of parameters be given that fulfill the constraints such that

$$\eta_{pi} = \theta_p - \beta_i - \mathbf{x}_p^T \boldsymbol{\gamma}_i = \tilde{\theta}_p - \tilde{\beta}_i - \mathbf{x}_p^T \tilde{\boldsymbol{\gamma}}_i$$

for all persons and items. From considering item I and person p one obtains by using $\beta_I = \tilde{\beta}_I = 0$ and $\boldsymbol{\gamma}_I^T = \tilde{\boldsymbol{\gamma}}_I^T = (0, \dots, 0)$ that $\theta_p = \tilde{\theta}_p$ holds. Therefore, one has $\beta_i + \mathbf{x}_p^T \boldsymbol{\gamma}_i = \tilde{\beta}_i + \mathbf{x}_p^T \tilde{\boldsymbol{\gamma}}_i$ for all p, i , which for item i can be written in matrix form as

$$\mathbf{X}(\beta_i, \boldsymbol{\gamma}_i)^T = \mathbf{X}(\tilde{\beta}_i, \tilde{\boldsymbol{\gamma}}_i)^T.$$

One can multiply on both sides of the equation with \mathbf{X}^T , and, since \mathbf{X} has full rank, with the inverse $(\mathbf{X}^T \mathbf{X})^{-1}$, obtaining $(\beta_i, \boldsymbol{\gamma}_i)^T = (\tilde{\beta}_i, \tilde{\boldsymbol{\gamma}}_i)^T$. Alternatively one can use the single value decomposition of \mathbf{X} .

C. Additional Results for the WC1994 Data in Chapter 9

































			Round of 16	Quarter finals	Semi finals	Final	World Champion	Oddset
1.		GER	86.1	68.1	52.3	32.8	20.5	14.2
2.		ESP	91.3	64.1	47.5	31.7	19.5	10.9
3.		BRA	93.0	64.9	48.2	30.8	19.1	20.3
4.		POR	73.3	51.1	35.1	18.7	9.3	2.4
5.		URU	71.3	50.7	22.5	11.5	5.1	2.8
6.		BEL	82.8	36.9	22.4	10.2	4.3	5.9
7.		ITA	67.2	46.3	19.5	9.4	4.0	3.5
8.		SUI	72.3	45.6	19.7	8.5	3.5	0.7
9.		ARG	77.6	44.5	18.9	7.8	3.1	14.2
10.		CRO	64.9	26.2	13.8	6.0	2.1	0.7
11.		FRA	62.2	35.4	13.7	5.3	1.9	3.5
12.		COL	76.3	33.4	10.9	4.1	1.3	3.9
13.		ENG	47.3	28.1	9.5	3.7	1.3	3.5
14.		CHI	50.1	18.0	8.6	3.3	1.0	2.0
15.		NED	44.9	15.1	6.9	2.5	0.7	3.5
16.		BIH	56.6	25.2	7.9	2.4	0.7	0.5
17.		ALG	49.3	13.2	5.6	1.6	0.4	0.1
18.		CIV	61.3	21.4	5.5	1.7	0.4	0.7
19.		USA	23.2	10.7	4.8	1.5	0.4	0.7
20.		ECU	38.8	17.3	4.8	1.3	0.3	0.7
21.		NGA	39.3	14.2	3.4	0.8	0.2	0.4
22.		RUS	42.7	9.0	3.5	0.8	0.2	1.2
23.		GHA	17.4	7.2	2.9	0.7	0.2	0.7
24.		MEX	28.0	6.9	2.4	0.7	0.2	0.7
25.		JPN	43.0	11.5	2.2	0.5	0.1	0.5
26.		HON	26.6	9.8	2.2	0.5	0.1	0.1
27.		IRN	26.4	7.9	1.6	0.3	0.1	0.1
28.		KOR	25.2	3.8	1.1	0.2	0.0	0.2
29.		CRC	14.2	5.6	1.0	0.2	0.0	0.1
30.		CMR	14.0	2.3	0.6	0.1	0.0	0.2
31.		AUS	13.7	2.4	0.6	0.1	0.0	0.2
32.		GRE	19.4	3.1	0.3	0.1	0.0	0.7

Table C.1.: Estimated probabilities (in %) for reaching the different stages in the FIFA World Cup 2014 for all 32 teams based on 100,000 simulation runs of the FIFA World Cup 2014 and based on the estimates of the WC1994 data together with winning probabilities based on the ODDSET odds.

































			Round of 16	Quarter finals	Semi finals	Final	World Champion
1.		GER	86.1	81.4	68.4	53.2	73.9
2.		ARG	77.6	48.4	47.5	54.8	26.1
3.		BRA	93.0	76.6	73.3	46.8	0.0
4.		NED	44.9	66.0	67.0	45.2	0.0
5.		BEL	82.8	65.7	52.5	0.0	0.0
6.		CRC	14.2	68.0	33.0	0.0	0.0
7.		FRA	62.2	68.8	31.6	0.0	0.0
8.		COL	76.3	41.6	26.7	0.0	0.0
9.		URU	71.3	58.4	0.0	0.0	0.0
10.		SUI	72.3	51.6	0.0	0.0	0.0
11.		USA	23.2	34.3	0.0	0.0	0.0
12.		MEX	28.0	34.0	0.0	0.0	0.0
13.		GRE	19.4	32.0	0.0	0.0	0.0
14.		NGA	39.3	31.2	0.0	0.0	0.0
15.		CHI	50.1	23.4	0.0	0.0	0.0
16.		ALG	49.3	18.6	0.0	0.0	0.0
17.		ESP	91.3	0.0	0.0	0.0	0.0
18.		POR	73.3	0.0	0.0	0.0	0.0
19.		ITA	67.2	0.0	0.0	0.0	0.0
20.		CRO	64.9	0.0	0.0	0.0	0.0
21.		CIV	61.3	0.0	0.0	0.0	0.0
22.		BIH	56.6	0.0	0.0	0.0	0.0
23.		ENG	47.3	0.0	0.0	0.0	0.0
24.		JPN	43.0	0.0	0.0	0.0	0.0
25.		RUS	42.7	0.0	0.0	0.0	0.0
26.		ECU	38.8	0.0	0.0	0.0	0.0
27.		HON	26.6	0.0	0.0	0.0	0.0
28.		IRN	26.4	0.0	0.0	0.0	0.0
29.		KOR	25.2	0.0	0.0	0.0	0.0
30.		GHA	17.4	0.0	0.0	0.0	0.0
31.		CMR	14.0	0.0	0.0	0.0	0.0
32.		AUS	13.7	0.0	0.0	0.0	0.0

Table C.2.: Estimated (adapted) probabilities (in %) for reaching the next stages in the FIFA World Cup 2014 for all 32 teams based on 100,000 simulation runs of the FIFA World Cup 2014. After each round, the data set (WC1994) is extended with by the matches already played and the model is refitted. Only actual matches from the World Cup are simulated.


Group A 43%	Group B 33%	Group C 24%	Group D 22%
1.  BRA	1.  ESP	1.  COL	1.  URU
2.  CRO	2.  CHI	2.  CIV	2.  ITA
 MEX	 NED	 JPN	 ENG
 CMR	 AUS	 GRE	 CRC
Group E 22%	Group F 24%	Group G 36%	Group H 24%
1.  SUI	1.  ARG	1.  GER	1.  BEL
2.  FRA	2.  BIH	2.  POR	2.  ALG
 ECU	 NGA	 GHA	 RUS
 HON	 IRN	 USA	 KOR

Table C.3.: Most probable final group standings together with the corresponding probabilities for the FIFA World Cup 2014 based on 100,000 simulation runs and on the estimates of the WC1994 data.

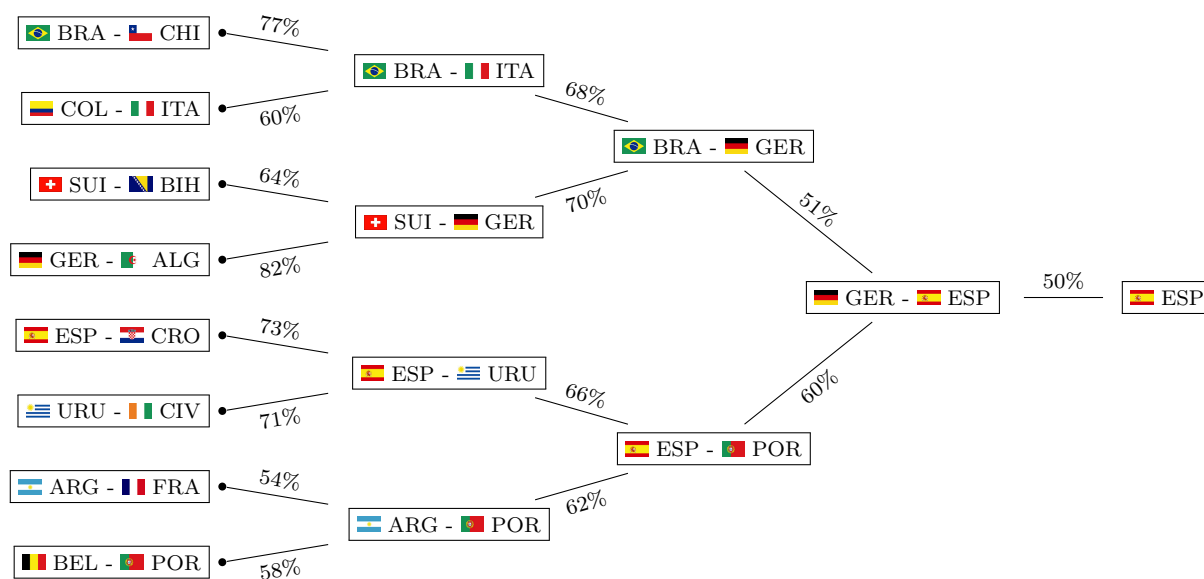


Figure C.1.: Most probable course of the knockout stage together with corresponding probabilities for the FIFA World Cup 2014 based on 100,000 simulation runs and on the estimates of the WC1994 data.

References

- Agresti, A. (1992). Analysis of ordinal paired comparison data. *Applied Statistics* 41(2), 287–297.
- Agresti, A. (2002). *Categorical Data Analysis*. New York: Wiley.
- Agresti, A. (2009). *Analysis of Ordinal Categorical Data, 2nd Edition*. New York: Wiley.
- Akaike, H. (1973). Information theory and the extension of the maximum likelihood principle. In B. Petrov and F. Caski (Eds.), *Second International Symposium on Information Theory*, Budapest. Akademia Kiado.
- Amthauer, R., B. Brocke, D. Liepmann, and A. Beauducel (1973). Intelligenz-Struktur-Test (IST 70). *Göttingen: Hogrefe*.
- Amthauer, R., B. Brocke, D. Liepmann, and A. Beauducel (1999). Intelligenz-Struktur-Test 2000 (IST 2000). *Göttingen: Hogrefe*.
- Amthauer, R., B. Brocke, D. Liepmann, and A. Beauducel (2001). Intelligenz-Struktur-Test 2000 R (IST 2000 R). *Göttingen: Hogrefe*.
- Andersen, E. (1973a). A goodness of fit test for the Rasch model. *Psychometrika* 38, 123–140. 10.1007/BF02291180.
- Andersen, E. B. (1973b). *Conditional Inference and Models for Measuring*. Copenhagen: Metalhygiejnisk Forlag.
- Andersen, E. B. (1980). *Discrete Statistical Models with Social Science Applications*. Amsterdam: North-Holland.
- Anderson, E. (1957). A semigraphical method for the analysis of complex problems. *Proceedings of the National Academy of Sciences of the United States of America* 43(10), 923.
- Archer, K. J. (2014a). *glmnetcr: Fit a penalized constrained continuation ratio model for predicting an ordinal response*. R package version 1.0.2.

- Archer, K. J. (2014b). *glmpathcr: Fit a penalized continuation ratio model for predicting an ordinal response*. R package version 1.0.3.
- Archer, K. J. and A. A. A. Williams (2012). L 1 penalized continuation ratio models for ordinal response prediction using high-dimensional datasets. *Statistics in Medicine* 31(14), 1464–1474.
- Birnbaum, A. (1968). Some latent trait models and their uses in inferring an examinee's ability. In F. F. M. Lord and M. Novick (Eds.), *Statistical theories of mental test score*, pp. 397–479. Reading, MA: Addison- Wesley.
- Böckenholt, U. and W. R. Dillon (1997a). Modeling within-subject dependencies in ordinal paired comparison data. *Psychometrika* 62(3), 411–434.
- Böckenholt, U. and W. R. Dillon (1997b). Some new methods for an old problem: Modeling preference changes and competitive market structures in pretest market data. *Journal of Marketing Research* 34(1), 130–142.
- Bondell, H. D., A. Krishna, and S. K. Ghosh (2010). Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics* 66(4), 1069–1077.
- Bondell, H. D. and B. J. Reich (2009). Simultaneous factor selection and collapsing levels in anova. *Biometrics* 65, 169–177.
- Boulesteix, A.-L. and T. Hothorn (2010). Testing the additional predictive value of high-dimensional molecular data. *BMC Bioinformatics* 11, 1–11.
- Bradley, R. A. (1976). Science, statistics, and paired comparison. *Biometrics* 32, 213–232.
- Bradley, R. A. (1984). Paired comparisons: Some basic procedures and examples. In P. Krishnaiah and P. R. Sen (Eds.), *Handbook of Statistics*, Volume 4, pp. 299–326. Elsevier.
- Bradley, R. A. and M. E. Terry (1952). Rank analysis of incomplete block designs, I: The method of pair comparisons. *Biometrika* 39, 324–345.
- Brant, R. (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics* 46, 1171–1178.
- Breiman, L. (2001). Random forests. *Machine Learning* 45, 5–32.
- Breiman, L., J. H. Friedman, R. A. Olshen, and J. C. Stone (1984). *Classification and Regression Trees*. Monterey, CA: Wadsworth.
- Brocke, B., A. Beauducel, and K. Tasche (1998). Der Intelligenz-Struktur-Test: Analysen zur theoretischen Grundlage und technischen Güte. *Diagnostica* 44, 84–99.

- Brüderl, J., P. Preisendörfer, and R. Ziegler (1992). Survival chances of newly founded business organizations. *American Sociological Review* 57, 227–242.
- Bühlmann, P. (2006). Boosting for high-dimensional linear models. *Annals of Statistics* 34, 559–583.
- Bühlmann, P. and T. Hothorn (2007a). Boosting algorithms: regularization, prediction and model fitting. *Statistical Science* 22, 477–505.
- Bühlmann, P. and T. Hothorn (2007b). Rejoinder: Boosting algorithms: regularization, prediction and model fitting. *Statistical Science* 22, 516–522.
- Bühlmann, P. and B. Yu (2003). Boosting with the L2 loss: Regression and classification. *Journal of the American Statistical Association* 98, 324–339.
- Bühner, M., M. Ziegler, S. Krumm, and L. Schmidt-Atzert (2006). Ist der IST 2000 R Rasch-skalierbar? *Diagnostica* 52(3), 119–130.
- Buja, A., T. Hastie, and R. Tibshirani (1989). Linear smoothers and additive models. *Annals of Statistics* 17, 453–510.
- Candell, G. L. and F. Drasgow (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement* 12(3), 253–260.
- Casalicchio, G. (2013). *ordBTL: Modelling comparison data with ordinal response*. R package version 0.7.
- Casalicchio, G., G. Tutz, and G. Schauburger (2015). Subject-specific Bradley–Terry–Luce models with implicit variable selection. *Statistical Modelling*, published online.
- Cattelan, M. (2012). Models for paired comparison data: A review with emphasis on dependent data. *Statistical Science* 27(3), 412–433.
- Cattelan, M., C. Varin, and D. Firth (2013). Dynamic Bradley-Terry modelling of sports tournaments. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 62(1), 135–150.
- Chen, C., W. Härdle, and A. Unwin (2008). *Handbook of data visualization*. Springer Handbooks of Computational Statistics. Springer.
- Chernoff, H. (1973). The use of faces to represent points in k-dimensional space graphically. *Journal of the American Statistical Association* 68(342), pp. 361–368.

- Clauser, B., K. Mazor, and R. K. Hambleton (1993). The effects of purification of matching criterion on the identification of dif using the Mantel-Haenszel procedure. *Applied Measurement in Education* 6(4), 269–279.
- Cleveland, W. S. (1985). *The elements of graphing data*. Belmont, CA, USA: Wadsworth Publ. Co.
- Cox, C. (1995). Location-scale cumulative odds models for ordinal data: A generalized non-linear model approach. *Statistics in Medicine* 14, 1191–1203.
- David, H. A. (1988). *The method of paired comparisons, 2nd ed.* Griffin's Statistical Monographs & Courses 41, Griffin, London.
- Davidson, R. (1970). On extending the Bradley-Terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association* 65, 317–328.
- Dittrich, R., B. Francis, R. Hatzinger, and W. Katzenbeisser (2007). A paired comparison approach for the analysis of sets of Likert-scale responses. *Statistical Modelling* 7(1), 3–28.
- Dittrich, R., R. Hatzinger, and W. Katzenbeisser (1998). Modelling the effect of subject-specific covariates in paired comparison studies with an application to university rankings. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 47(4), 511–525.
- Dittrich, R., R. Hatzinger, and W. Katzenbeisser (2004). A log-linear approach for modelling ordinal paired comparison data on motives to start a phd programme. *Statistical Modelling* 4(3), 181–193.
- Dittrich, R., W. Katzenbeisser, and H. Reisinger (2000). The analysis of rank ordered preference data based on Bradley-Terry type models. *OR-Spektrum* 22(1), 117–134.
- Dixon, M. J. and S. G. Coles (1997). Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 46(2), 265–280.
- Dobson, S. and J. Goddard (2011). *The economics of football*. Cambridge University Press, Cambridge.
- Du Toit, S. H. C., A. G. W. Steyn, and R. H. Stumpf (1986). *Graphical exploratory data analysis*. New York, NY, USA: Springer-Verlag New York, Inc.
- Dyte, D. and S. R. Clarke (2000). A ratings based Poisson model for World Cup soccer simulation. *Journal of the Operational Research Society* 51 (8), 993–998.
- Eddelbuettel, D. (2013). *Seamless R and C++ integration with Rcpp*. Springer.

- Eddelbuettel, D., R. François, J. Allaire, J. Chambers, D. Bates, and K. Ushey (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software* 40(8), 1–18.
- Eddelbuettel, D. and C. Sanderson (2014). Rcpparmadillo: Accelerating R with high-performance C++ linear algebra. *Computational Statistics and Data Analysis* 71, 1054–1063.
- Eilers, P. H. C. and B. D. Marx (1996). Flexible smoothing with B-splines and Penalties. *Statistical Science* 11, 89–121.
- Elo, A. E. (2008). *The Rating of Chess Players, Past and Present*. San Rafael: Ishi Press.
- Eugster, M. J. A., J. Gertheiss, and S. Kaiser (2011). Having the second leg at home - advantage in the UEFA Champions League knockout phase? *Journal of Quantitative Analysis in Sports* 7(1), Article 6.
- Fahrmeir, L. and G. Tutz (1994). Dynamic stochastic models for time-dependent ordered paired comparison systems. *Journal of the American Statistical Association* 89, 1438–1449.
- Fahrmeir, L. and G. Tutz (2001). *Multivariate Statistical Modelling based on Generalized Linear Models*. New York: Springer.
- Fan, J. and R. Li (2001). Variable selection via nonconcave penalize likelihood and its oracle properties. *Journal of the American Statistical Association* 96, 1348–1360.
- Fidalgo, A., G. J. Mellenbergh, and J. Muñoz (2000). Effects of amount of DIF, test length, and purification type on robustness and power of Mantel-Haenszel procedures. *Methods of Psychological Research Online* 5(3), 43–53.
- Firth, D. and R. De Menezes (2004). Quasi-variances. *Biometrika* 91, 65.
- Fischer, G. H. and I. W. Molenaar (1995). *Rasch models: Foundations, recent developments, and applications*. Springer Science & Business Media.
- Forrest, D. and R. Simmons (2000). Forecasting sport: the behaviour and performance of football tipsters. *International Journal of Forecasting* 16(3), 317 – 331.
- Fox, J. (2003). Effect displays in R for generalised linear models. *Journal of Statistical Software* 8(15), 1–27.
- Fox, J. and R. Andersen (2006). Effect displays for multinomial and proportional-odds logit models. *Sociological Methodology* 36(1), 225–255.

- Fox, J. and J. Hong (2009). Effect displays in R for multinomial and proportional-odds logit models: Extensions to the effects package. *Journal of Statistical Software* 32(1), 1–24.
- Francis, B., R. Dittrich, and R. Hatzinger (2010). Modeling heterogeneity in ranked responses by nonparametric maximum likelihood: How do europeans get their scientific knowledge? *The Annals of Applied Statistics* 4(4), 2181–2202.
- Francis, B., R. Dittrich, R. Hatzinger, and R. Penn (2002). Analysing partial ranks by using smoothed paired comparison methods: an investigation of value orientation in europe. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 51, 319–336.
- Freund, Y. and R. E. Schapire (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences* 55, 119–139.
- Freund, Y., R. E. Schapire, et al. (1996). Experiments with a new boosting algorithm. In *ICML*, Volume 96, pp. 148–156.
- Friedman, J. H., T. Hastie, and R. Tibshirani (2000). Additive logistic regression: A statistical view of boosting. *Annals of Statistics* 28, 337–407.
- Friedman, J. H., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 33(1), 1–22.
- Friendly, M. (1994). Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association* 89(425), 190–200.
- Fu, W. J. (1998). Penalized regression: the bridge versus the lasso. *Journal of Computational and Graphical Statistics* 7, 397–416.
- Gertheiss, J. and G. Tutz (2010). Sparse modeling of categorical explanatory variables. *Annals of Applied Statistics* 4, 2150–2180.
- Glickman, M. E. and H. S. Stern (1998). A state-space model for national football league scores. *Journal of the American Statistical Association* 93(441), 25–35.
- Gnanadesikan, R. (1977). *Methods for statistical data analysis of multivariate observations*. Wiley series in probability and mathematical statistics. Wiley.
- Goldman-Sachs Global Investment Research (2014). The world cup and economics 2014. <http://www.goldmansachs.com/our-thinking/outlook/world-cup-and-economics-2014-folder/world-cup-economics-report.pdf>.
- Gonçalves, F., D. Gamerman, and T. Soares (2013). Simultaneous multifactor DIF analysis and detection in item response theory. *Computational Statistics & Data Analysis* 59, 144 – 160.

- Groll, A. and J. Abedieh (2013). Spain retains its title and sets a new record - generalized linear mixed models on European football championships. *Journal of Quantitative Analysis in Sports* 9(1), 51–66.
- Groll, A., G. Schauburger, and G. Tutz (2014). Brazil or Germany - who will win the trophy? prediction of the FIFA World Cup 2014 based on team-specific regularized poisson regression. Technical Report 166, Department of Statistics, Ludwig-Maximilians-Universität München, Germany.
- Groll, A., G. Schauburger, and G. Tutz (2015). Prediction of major international soccer tournaments based on team-specific regularized poisson regression: An application to the fifa world cup 2014. *Journal of Quantitative Analysis in Sports* 11(2), 97–115.
- Groll, A. and G. Tutz (2014). Variable selection for generalized linear mixed models by L_1 -penalized estimation. *Statistics and Computing* 24(2), 137–154.
- Hastie, T. (2007). Comment: Boosting algorithms: regularization, prediction and model fitting. *Statistical Science* 22, 513–515.
- Hastie, T., R. Tibshirani, and J. H. Friedman (2009). *The Elements of Statistical Learning (Second Edition)*. New York: Springer-Verlag.
- Hatzinger, R. (1989). The Rasch model, some extensions and their relation to the class of generalized linear models. In A. Decarli, B. Francis, R. Gilchrist, and G. Seeber (Eds.), *Statistical Modelling*, Volume 57 of *Lecture Notes in Statistics*, pp. 172–179. Springer New York.
- Hatzinger, R. and R. Dittrich (2012). pfmmod: An R package for modeling preferences based on paired comparisons, rankings, or ratings. *Journal of Statistical Software* 48(10), 1–31.
- Hoerl, A. E. and R. W. Kennard (1970). Ridge regression: Bias estimation for nonorthogonal problems. *Technometrics* 12, 55–67.
- Hofner, B., T. Hothorn, T. Kneib, and M. Schmid (2011). A framework for unbiased model selection based on boosting. *Journal of Computational and Graphical Statistics* 20(4), 956–971.
- Holland, P. W. and D. T. Thayer (1988). Differential item performance and the Mantel-Haenszel procedure. *Test validity*, 129–145.
- Holland, P. W. and H. Wainer (2012). *Differential item functioning*. Hoboken, NJ: Taylor and Francis.

- Hothorn, T., P. Buehlmann, T. Kneib, M. Schmid, and B. Hofner (2013). *mboost: Model-Based Boosting*. R package version 2.2-3.
- Karlis, D. and I. Ntzoufras (2003). Analysis of sports data by using bivariate poisson models. *The Statistician* 52, 381–393.
- Karlis, D. and I. Ntzoufras (2011). Robust fitting of football prediction models. *IMA Journal of Management Mathematics* 22(2), 171–182.
- Kastellec, J. and E. Leoni (2007). Using graphs instead of tables in political science. *Perspectives on Politics* 5(4), 755–771.
- Kauermann, G., G. Tutz, and J. Brüderl (2005). The survival of newly founded firms: A case-study into varying-coefficient models. *Journal of the Royal Statistical Society A* 168, 145–158.
- Kelderman, H. (1984). Loglinear Rasch model tests. *Psychometrika* 49(2), 223–245.
- Kim, S.-H., A. S. Cohen, and T.-H. Park (1995). Detection of differential item functioning in multiple groups. *Journal of Educational Measurement* 32(3), 261–276.
- Knight, K. and W. Fu (2000, 10). Asymptotics for lasso-type estimators. *Ann. Statist.* 28(5), 1356–1378.
- Knorr-Held, L. (2000). Dynamic rating of sports teams. *Journal of the Royal Statistical Society: Series D (The Statistician)* 49(2), 261–276.
- Koopman, S. J. and R. Lit (2015). A dynamic bivariate poisson model for analysing and forecasting match results in the English Premier League. *Journal of the Royal Statistical Society A* 178, 167–186.
- Kopf, J., A. Zeileis, and C. Strobl (2015). Anchor selection strategies for dif analysis: Review, assessment, and new approaches. *Educational and Psychological Measurement* 75(1), 22–56.
- Kuk, A. Y. C. (1995). Modelling paired comparison data with large numbers of draws and large variability of draw percentages among players. *Journal of the Royal Statistical Society. Series D (The Statistician)* 44(4), pp. 523–528.
- LeCessie (1992). Ridge estimators in logistic regression. *Applied Statistics* 41(1), 191–201.
- Lee, A. J. (1997). Modeling scores in the premier league: is manchester united really the best? *Chance* 10, 15–19.

- Leitner, C., A. Zeileis, and K. Hornik (2010a). Forecasting sports tournaments by ratings of (prob)abilities: A comparison for the EURO 2008. *International Journal of Forecasting* 26(3), 471–481.
- Leitner, C., A. Zeileis, and K. Hornik (2010b). Forecasting the winner of the FIFA World Cup 2010. Research Report Series, 100 Report 100, Department of Statistics and Mathematics, University of Vienna.
- Lloyd's (2014). Fifa world cup: How much are those legs worth? <http://www.lloyds.com/news-and-insight/news-and-features/market-news/industry-news-2014/fifa-world-cup-how-much-are-those-leg-worth>.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Routledge.
- Luce, R. D. (1959). *Individual Choice Behaviour*. New York: Wiley.
- Magis, D., S. Beland, and G. Raiche (2013). *difR: Collection of methods to detect dichotomous differential item functioning (DIF) in psychometrics*. R package version 4.4.
- Magis, D., S. Bèland, F. Tuerlinckx, and P. Boeck (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods* 42(3), 847–862.
- Magis, D., G. Raîche, S. Béland, and P. Gérard (2011). A generalized logistic regression procedure to detect differential item functioning among multiple groups. *International Journal of Testing* 11(4), 365–386.
- Magis, D., F. Tuerlinckx, and P. De Boeck (2015). Detection of differential item functioning using the lasso approach. *Journal of Educational and Behavioral Statistics* 40(2), 111–135.
- Maher, M. J. (1982). Modelling association football scores. *Statistica Neerlandica* 36, 109–118.
- Mair, P. and R. Hatzinger (2007). Extended Rasch modeling: The erm package for the application of irt models in R. *Journal of Statistical Software* 20(9), 1–20.
- Mair, P., R. Hatzinger, and M. J. Maier (2012). *eRm: Extended Rasch Modeling*. R package version 0.15-0.
- Mantel, N. and W. Haenszel (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* 22(4), 719–748.
- Masarotto, G. and C. Varin (2012). The ranking lasso and its application to sport tournaments. *The Annals of Applied Statistics* 6(4), 1949–1970.

- Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika* 47(2), 149–174.
- McCullagh, P. (1980). Regression model for ordinal data (with discussion). *Journal of the Royal Statistical Society B* 42, 109–127.
- McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models (Second Edition)*. New York: Chapman & Hall.
- McHale, I. G. and P. A. Scarf (2006). Forecasting international soccer match results using bivariate discrete distributions. Technical Report 322, Working paper, Salford Business School.
- McHale, I. G. and P. A. Scarf (2011). Modelling the dependence of goals scored by opposing teams in international soccer matches. *Statistical Modelling* 41(3), 219–236.
- Meier, L. (2009). *grplasso: Fitting user specified models with Group Lasso penalty*. R package version 0.4-2.
- Meier, L., S. van de Geer, and P. Bühlmann (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society, Series B* 70, 53–71.
- Meinshausen, N. and P. Bühlmann (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72(4), 417–473.
- Merkle, E. C. and A. Zeileis (2013). Tests of measurement invariance without subgroups: A generalization of classical methods. *Psychometrika* 78, 59–82.
- Meyer, D., A. Zeileis, and K. Hornik (2008). Visualizing contingency tables. In C. Chen, W. Härdle, and A. Unwin (Eds.), *Handbook of Data Visualization*, Springer Handbooks of Computational Statistics, pp. 589–616. Springer Berlin Heidelberg.
- Millsap, R. and H. Everson (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement* 17(4), 297–334.
- Ni, X., D. Zhang, and H. H. Zhang (2010). Variable selection for semiparametric mixed models in longitudinal studies. *Biometrics* 66, 79–88.
- Nyquist, H. (1991). Restricted estimation of generalized linear models. *Applied Statistics* 40, 133–141.
- Oelker, M.-R. (2015). *gvcm.cat: Regularized Categorical Effects/Categorical Effect Modifiers/Continuous/Smooth Effects in GLMs*. R package version 1.9.

- Oelker, M.-R., J. Gertheiss, and G. Tutz (2014). Regularization and model selection with categorical predictors and effect modifiers in generalized linear models. *Statistical Modelling* 14(2), 157–177.
- Oelker, M.-R. and G. Tutz (2015). A uniform framework for the combination of penalties in generalized structured models. *Advances in Data Analysis and Classification*, published online.
- Osborne, M., B. Presnell, and B. Turlach (2000). On the lasso and its dual. *Journal of Computational and Graphical Statistics* 9(2), 319–337.
- Osterlind, S. and H. Everson (2009). *Differential item functioning*, Volume 161. Sage Publications, Inc.
- Paek, I. and M. Wilson (2011). Formulating the Rasch differential item functioning model under the marginal maximum likelihood estimation context and its comparison with mantel–haenszel procedure in short test and small sample conditions. *Educational and Psychological Measurement* 71(6), 1023–1046.
- Park, M. Y. and T. Hastie (2007). An l1 regularization-path algorithm for generalized linear models. *Journal of the Royal Statistical Society B* 69, 659–677.
- Penfield, R. D. (2001). Assessing differential item functioning among multiple groups: A comparison of three mantel-haenszel procedures. *Applied Measurement in Education* 14(3), 235–259.
- Peterson, B. and F. E. Harrell (1990). Partial proportional odds models for ordinal response variables. *Applied Statistics* 39, 205–217.
- Plass, J., P. Fink, N. Schöning, and T. Augustin (2015). Statistical modelling in surveys without neglecting "the undecided": Multinomial logistic regression models and imprecise classification trees under ontic data imprecision - extended version. Technical Report 179, Department of Statistics, Ludwig-Maximilians-Universität München, Germany.
- R Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika* 53(4), 495–502.
- Rao, P. and L. Kupper (1967). Ties in paired-comparison experiments: A generalization of the Bradley-Terry model. *Journal of the American Statistical Association* 62, 194–204.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research.

- Rattinger, H., S. Roßteutscher, R. Schmitt-Beck, B. Weßels, and C. Wolf (2014). Pre-election cross section (GLES 2013). *GESIS Data Archive, Cologne ZA5700 Data file Version 2.0.0*.
- Rogers, H. J. (2005). *Differential Item Functioning*. John Wiley & Sons, Ltd.
- Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement* 14(3), 271–282.
- Rue, H. and O. Salvesen (2000). Prediction and retrospective analysis of soccer matches in a league. *Journal of the Royal Statistical Society: Series D (The Statistician)* 49(3), 399–418.
- Samejima, F. (1997). Graded response model. In W. van der Linden and R. Hambleton (Eds.), *Handbook of Modern Item Response Theory*, pp. 85–100. Springer New York.
- Schauberger, G. (2014a). *DIFlasso: A penalty approach to Differential Item Functioning in Rasch Models*. R package version 1.0-1.
- Schauberger, G. (2014b). *EffectStars: Visualization of Categorical Response Models*. R package version 1.5.
- Schauberger, G. (2015a). *BTLasso: Modelling Heterogeneity in Paired Comparison Data*. R package version 0.1-2.
- Schauberger, G. (2015b). *DIFboost: Detection of Differential Item Functioning (DIF) in Rasch Models by Boosting Techniques*. R package version 0.1.
- Schauberger, G. and G. Tutz (2012). Effect stars for categorical response models. In A. Komarek and S. Nagy (Eds.), *Proceedings of the 27th International Workshop on Statistical Modelling*, Volume 2, pp. 729–734. Statistical Modelling Society.
- Schauberger, G. and G. Tutz (2013). DIF-LASSO: Differential item functioning in rasch models. In V. M. Muggeo, V. Capursi, G. Boscaino, and G. Lovison (Eds.), *Proceedings of the 28th International Workshop on Statistical Modelling*, Volume 1, pp. 375–379. Statistical Modelling Society.
- Schauberger, G. and G. Tutz (2014). DIFboost: A boosting method for the detection of differential item functioning. In T. Kneib, F. Sobotka, J. Fahrenholz, and H. Irmer (Eds.), *Proceedings of the 29th International Workshop on Statistical Modelling*, Volume 1, pp. 319–323. Statistical Modelling Society.
- Schauberger, G. and G. Tutz (2015a). BTL-Lasso - a penalty approach to heterogeneity in paired comparison data. In H. Friedl and H. Wagner (Eds.), *Proceedings of the 30th International Workshop on Statistical Modelling*, Volume 1, pp. 336–341. Statistical Modelling Society.

- Schauberger, G. and G. Tutz (2015b). Detection of differential item functioning in Rasch models by boosting techniques. *British Journal of Mathematical and Statistical Psychology*, published online.
- Schauberger, G. and G. Tutz (2015c). Modelling heterogeneity in paired comparison data - an L1 penalty approach with an application to party preference data. Technical Report 183, Department of Statistics, Ludwig-Maximilians-Universität München, Germany.
- Schmidt-Atzert, L. (2002). Intelligenz-Struktur-Test 2000 R (Testrezension). *Zeitschrift für Personalpsychologie* 1, 50–56.
- Schmidt-Atzert, L., W. Hommers, and M. Heß (1995). Der I-S-T 70. Eine Analyse und Neubewertung. *Diagnostica* 41, 108–130.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* 6, 461–464.
- Segerstedt, B. (1992). On ordinary ridge regression in generalized linear models. *Communications in Statistics – Theory and Methods* 21, 2227–2246.
- Siegel, J., E. Farrell, R. Goldwyn, and H. Friedman (1972). The surgical implications of physiologic patterns in myocardial infarction shock. *Surgery* 72(1), 126–141.
- Silver, N. (2014). It's Brazil's World Cup to Lose. <http://fivethirtyeight.com/features/its-brazils-world-cup-to-lose/>.
- Soares, T., F. Gonçalves, and D. Gamerman (2009). An integrated bayesian model for dif analysis. *Journal of Educational and Behavioral Statistics* 34(3), 348–377.
- Somes, G. W. (1986). The generalized Mantel–Haenszel statistic. *The American Statistician* 40(2), 106–108.
- Springall, A. (1973). Response surface fitting using a generalization of the Bradley-Terry paired comparison model. *Applied Statistics* 22, 59–68.
- Stoy, V., R. Frankenberger, D. Buhr, L. Haug, B. Springer, and J. Schmid (2010). Das Ganze ist mehr als die Summe seiner Lichtgestalten. Eine ganzheitliche Analyse der Erfolgchancen bei der Fußballweltmeisterschaft 2010. Working Paper 46, Eberhard Karls University, Tübingen, Germany.
- Strobl, C., J. Kopf, and A. Zeileis (2015). Rasch trees: A new method for detecting differential item functioning in the Rasch model. *Psychometrika* 80(2), 289–316.
- Strobl, C., J. Malley, and G. Tutz (2009). An Introduction to Recursive Partitioning: Rationale, Application and Characteristics of Classification and Regression Trees, Bagging and Random Forests. *Psychological Methods* 14, 323–348.

- Strobl, C., F. Wickelmaier, and A. Zeileis (2011). Accounting for individual differences in Bradley-Terry models by means of recursive partitioning. *Journal of Educational and Behavioral Statistics* 36(2), 135–153.
- Swaminathan, H. and H. J. Rogers (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement* 27(4), 361–370.
- Theus, M. and S. Lauer (1999). Visualizing loglinear models. *Journal of Computational and Graphical Statistics* 8(3), 396–412.
- Thissen, D., L. Steinberg, and H. Wainer (1993). Detection of differential item functioning using the parameters of item response models. In P. Holland and H. Wainer (Eds.), *Differential item functioning*, pp. 67–113. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B* 58, 267–288.
- Townsend, Z., J. Buckley, M. Harada, and M. Scott (2013). The choice between fixed and random effects. In M. Scott, J. Simonoff, and B. Marx (Eds.), *The SAGE Handbook of Multilevel Modeling*. SAGE.
- Tukey, J. (1977). *Exploratory Data Analysis*. Reading, Pennsylvania: Addison Wesley.
- Turner, H. and D. Firth (2012). Bradley-Terry models in R: The BradleyTerry2 package. *Journal of Statistical Software* 48(9), 1–21.
- Tutz, G. (1986). Bradley-Terry-Luce models with an ordered response. *Journal of Mathematical Psychology* 30, 306–316.
- Tutz, G. (1989). *Latent Trait-Modelle für ordinale Beobachtungen: die statistische und messtheoretische Analyse von Paarvergleichsdaten*, Volume 30. Springer-Verlag.
- Tutz, G. (2012). *Regression for Categorical Data*. Cambridge University Press.
- Tutz, G. and H. Binder (2006). Generalized additive modeling with implicit variable selection by likelihood-based boosting. *Biometrics* 62, 961–971.
- Tutz, G. and G. Schauburger (2012a). A penalty approach to differential item functioning in Rasch models. Technical Report 134, Department of Statistics, Ludwig-Maximilians-Universität München, Germany.
- Tutz, G. and G. Schauburger (2012b). Visualization of categorical response models - from data glyphs to parameter glyphs. Technical Report 117, Department of Statistics, Ludwig-Maximilians-Universität München, Germany.

- Tutz, G. and G. Schauberger (2013). Visualization of categorical response models: From data glyphs to parameter glyphs. *Journal of Computational and Graphical Statistics* 22(1), 156–177.
- Tutz, G. and G. Schauberger (2014). Extended ordered paired comparison models with application to football data from German Bundesliga. Technical Report 151, Department of Statistics, Ludwig-Maximilians-Universität München, Germany.
- Tutz, G. and G. Schauberger (2015a). Extended ordered paired comparison models with application to football data from German Bundesliga. *AStA Advances in Statistical Analysis* 99(2), 209–227.
- Tutz, G. and G. Schauberger (2015b). A penalty approach to differential item functioning in Rasch models. *Psychometrika* 80(1), 21–43.
- Van den Noortgate, W. and P. De Boeck (2005). Assessing and explaining differential item functioning using logistic mixed models. *Journal of Educational and Behavioral Statistics* 30(4), 443–464.
- Wang, W.-C. and Y.-H. Su (2004). Effects of average signed area between two item characteristic curves and test purification procedures on the dif detection via the Mantel-Haenszel method. *Applied Measurement in Education* 17(2), 113–144.
- Woods, C. M. (2009). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement* 33(1), 42–57.
- Yee, T. (2010). The VGAM package for categorical data analysis. *Journal of Statistical Software* 32(10), 1–34.
- Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society B* 68, 49–67.
- Zeileis, A., T. Hothorn, and K. Hornik (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics* 17(2), 492–514.
- Zeileis, A., C. Leitner, and K. Hornik (2012). History repeating: Spain beats Germany in the EURO 2012 final. Working paper, Faculty of Economics and Statistics, University of Innsbruck.
- Zeileis, A., C. Leitner, and K. Hornik (2014). Home Victory for Brazil in the 2014 FIFA World Cup. Working paper, Faculty of Economics and Statistics, University of Innsbruck.
- Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* 101(476), 1418–1429.

- Zou, H., T. Hastie, and R. Tibshirani (2007). On the “degrees of freedom” of the lasso. *The Annals of Statistics* 35(5), 2173–2192.
- Zumbo, B. (1999). A handbook on the theory and methods of differential item functioning (dif). *Ottawa: National Defense Headquarters*.

Eidesstattliche Versicherung

(Siehe Promotionsordnung vom 12. Juli 2011, § 8 Abs. 2 Pkt. 5)

Hiermit erkläre ich an Eidesstatt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

München, den

Gunther Schauburger

